Università degli studi di Pisa

Facoltà di filologia, letteratura e linguistica

Corso di laurea magistrale in Informatica Umanistica

Tesi di laurea

Storia e struttura del Data Journalism

Relatori

Prof. Dino Pedreschi

Dr. Luca Pappalardo

Dott. Paolo Cintia

Candidato

Paolo Salvatore Locci

Anno accademico 2014/2015

SOMMARIO

ABSTRACT

(Italiano)

Gli obbiettivi di questa tesi sono di analizzare la nascita e lo sviluppo del data journalism a partire dalle inchieste giornalistiche che hanno determinato la sua evoluzione, analizzando il metodo di lavoro di tre premi Pulitzer, *Philip Meyer*, *Bill Dedman* e *Stephen K. Doig*. Esaminare quali sono i metodi di lavoro e gli strumenti più utilizzati dalle redazioni che sono più attente al data journalism, per arrivare alla creazione di un vero articolo di data journalism, "*La crisi economica e il declino del calcio italiano*", nel quale vengono messi in relazione i dati che riguardano la crisi economica e i dati che riguardano il declino del calcio Italiano, il quale dal 2010 non è stato all'altezza della propria tradizione calcistica. In questo periodo, in Italia, è stato registrato un vero e proprio crollo dal punto di vista dei risultati, da imputare ad un calo degli investimenti che non è stato riscontrato negli altri campionati europei, nei quali, a dispetto della crisi, gli investimenti sono aumentati.

(English)

The main goal of this work is to analyse origins and development of the Data Journalism from the journalistic investigations that have allowed his growth through the analysis of three Pulitzer prices, *Philip Meyer*, *Bill Dedman* and *Stephen K. Doig*. Analyse the methods of operation and the tools of the newsrooms more interested in the data journalism, to create a data journalism investigation "*The economic crisis and the Italian football decline*", in which I relate the data of the economic crisis and the data of the Italian football result. Since 2010 the results haven's been like past years, in this period there was a collapse of the results, caused by the reductions of investments which is not append in the others European championships, in which, despite to the crisis, the investment have increased.

INTRODUZIONE

"I fatti sono sacri, le opinioni sono libere"

Charles Prestwich Scott, storico giornalista britannico, scrisse questa frase in un saggio pubblicato nel 1921 per rappresentare gli ideali ai quali dovrebbe ispirarsi tutto il mondo del giornalismo, che si basa troppo su opinioni e poco sui fatti. Il data journalism è un giornalismo che si basa sui dati, intesi come raccolta di fatti e incarna a pieno la filosofia di C. P. Scott. Il data journalism ha l'obbiettivo di mettere i fatti al centro della notizia, i fatti vengono visti come un elemento indiscutibile dal quale partire per sviluppare la notizia, in maniera tale da conferirle maggiore precisione e obbiettività. Mentre il giornalismo tradizionale si limita a osservare e riportare i fatti secondo il punto di vista del giornalista, con il data journalism dobbiamo parlare di vere e proprie inchieste giornalistiche sulle quali i giornalisti indagano sui dati per arrivare alle conclusioni. Un altro elemento che contraddistingue il data journalism è la sua trasparenza, la sua filosofia di condivisione dei dati i quali possono essere utilizzati e ridistribuiti, in modo da permettere a chiunque di verificare la veridicità dei dati pubblicati, inoltre il suo prodotto finale non sono solo le notizie ma sono proprio i dati, che possono essere sviluppati nel tempo.

I principali obbiettivi di questa tesi sono tre:

- Analizzare la nascita e lo sviluppo del data journalism a partire dalla sue radici e dalle inchieste che ne hanno determinato l'evoluzione, analizzando il metodo di lavoro dei suoi pionieri: Philip Meyer che con l'applicazione del metodo scientifico è riuscito a individuare la reali cause delle rivolte avvenute a Detroit nel 1967, scaturite da parte della popolazione di colore; Bill Dedman che con una serie di articoli ha portato alla luce un caso di discriminazione razziale da parte di banche e istituti di credito, ad Atlanta, nel 1988; Stephen Doig che in occasione dell'uragano avvenuto nella città di Miami nel 1992 ha portato alla ribalta un caso di corruzione edilizia. I metodi di lavoro di questi giornalisti sono oggi un modello da seguire, un esempio del modo in cui è possibile usare i dati per avere una visione più profonda della vicende esaminate.

- Analizzare i metodi di lavoro delle redazioni che rappresentano i migliori esempi di data journalism e le iniziative, come il *Data journalism Awords* e il *The Editors Lab*, che mirano a incentivare l'uso e lo sviluppo di nuove tecnologie digitali e a favorire lo sviluppo di un giornalismo di alta qualità. Inoltre si analizza la filosofia di condivisione dettata dal progetto degli open data, il quale ha fatto sì che sempre più governi pubblicassero i propri dati e ha permesso un maggiore controllo sulla qualità dei dati pubblicati. Oltre a questo si mostrano in che modo i giornalisti sportivi, usano i dati per analizzare e prevedere l'andamento di alcuni eventi sportivi, in particolare i Playoff NBA e i mondiali di calcio.
- Sviluppare un'inchiesta giornalistica dal titolo *"La Crisi economica e il declino del calcio italiano"*, da pubblicare sul sito *bigdatatales*.com. Nell'inchiesta si esamina in che modo le principali squadre dei cinque campionati che tradizionalmente investono più soldi nel calcio, hanno reagito alla crisi economica che in Europa ha avuto il peggior momento nel 2009. Questa inchiesta giornalistica, che prende come punto di riferimento il valore medio delle squadre, evidenzia il fatto che le principali squadre italiane di calcio, dal momento della crisi, hanno deciso di tagliare i loro investimenti, a differenza delle squadre delle altre nazioni europee dove gli investimenti, a dispetto della crisi, hanno continuato ad aumentare. Inoltre l'inchiesta mostra in che modo i giocatori più valutati hanno progressivamente abbandonato il campionato italiano a favore degli altri campionati europei.

Il data journalism non si pone come un'alternativa al giornalismo tradizionale, ma come la sua evoluzione, nella quale si mescolano il classico fiuto per le notizie e il metodo scientifico. Il data journalism è l'applicazione del criterio di scrupolosità che un giornalista dovrebbe sempre avere nell'analizzare i fatti prima di pubblicarli. Il semplice fatto di utilizzare dei dati all'interno di un articolo non implica che si tratti di una inchiesta di data journalism, i dati devono essere scelti con cura secondo il criterio scientifico, tanto elogiato da Meyer, ed è proprio la fase di raccolta, selezione ed elaborazione dati la parte che contraddistingue il lavoro di un data journalist.

1 LE ORIGINI DEL DATA JOURNALISM

Il data journalism è un approccio giornalistico basato sui dati, a cavallo tra ricerca e inchiesta giornalistica. Spesso ci si riferisce a questa pratica giornalistica usando diversi nomi, da giornalismo di precisione, "Data-Driven Journalism" (Giornalismo guidato dai dati) e "Computer-Assisted Reported" (Giornalismo supportato dal computer) sino al termine più moderno "data journalism" (Giornalismo dei dati).

Il data journalism può essere definito come l'applicazione del metodo e del rigore scientifico al giornalismo, in questa pratica giornalistica i dati vengono scelti, selezionati ed elaborati secondo il criterio scientifico, che conferisce alla notizia maggiore obbiettività. Secondo la filosofia del data journalism i dati possono essere sia la fonte dalla quale si parte per creare le notizie oppure possono impersonare la notizia[1].

1.1 LE ORIGINI

Il concetto di giornalismo di precisione è più vecchio del termine stesso. Nonostante fosse già una pratica comune per molti giornalisti, esso diventa popolare solo dopo la pubblicazione del libro di Philip Meyer, *Precision Journalism: A Reporter's Introduction to Social Science Methods*, del 1973. Egli nella scelta del titolo prende spunto dalla parole pronunciate da Everette E. Dennis, nel 1971, in una conferenza tenutasi nell'Università dell'Oregon. Il termine viene usato soprattutto per distinguere quest'approccio dal "Nuovo Giornalismo", la cui espressione venne coniata dal Tom Wolfe nel 1973, indica un movimento culturale del giornalismo degli anni sessanta e settanta con uno stile di scrittura tra letteratura e giornalismo, possiede motivi tipici della letteratura, una delle forme più consolidate è il romanzo-reportage.

Negli anni settanta si sviluppa la rivalità tra questi due generi, due modi di vedere il giornalismo molto diversi e spesso inconciliabili. Lo scontro trova i massimi esponenti in Philip Meyer, Bill Dedman e Steve Doig per il giornalismo di precisione e Tom Wolfe e Gay

[1] Fonte : http://datajournalismhandbook.org/1.0/en/introduction_0.html

Talese per il giornalismo narrativo. Secondo la filosofia di Meyer, i giornalisti devono usare delle tecniche scientifiche, che permettano di aumentare la profondità e l'accuratezza delle proprie storie, a questo scopo i giornalisti devono confrontare il maggior numeri di dati dei quali riescono a disporre, in modo da conferire alle notizie maggiore precisione ed obbiettività. Sempre secondo Meyer, infatti, il giornalista non può avere a che fare con la dimensione del romanzo, che non è un genere che si addice al giornalismo, che deve essere fatto di verità controllate e non interpretate. Il giornalista viene paragonato a uno storico e a uno scienziato, deve acquisire competenze tali da permettergli di creare notizie giornalistiche fatte di ricostruzioni documentate. I fatti e le notizie esistono già per loro natura, il giornalista deve solo dare loro voce. Naturalmente le notizie non sono esenti dall'interpretazione personale, e non si può riportare un fatto in maniera completamente imparziale, ma la completezza dell'informazione non dipende dalla quantità di fatti raccolti in essa, ma dallo loro selezione obiettiva ed empirica. Sotto questo punto di vista il metodo scientifico è l'unico strumento valido per far fronte alle degenerazioni provocate dalle prospettive dell'intrattenimento, dalla pubblicità e dai soldi delle multinazionali. Tutto questo si contrappone alla filosofia del nuovo giornalismo, nel quale la frustrazione per l'irraggiungibilità dell'obbiettività aveva trasformato alcuni giornalisti in dei narratori, che sfruttavano gli espedienti del romanzo, nel quale si dà più importanza al modo in cui si narra la notizia che alla notizia in se.

1.2 LE INCHIESTE PIÙ CELEBRI DEL DATA JOURNALISM

I primi esempi di inchieste di giornalismo di precisione risalgono alla fine degli anni sessanta. Le tematiche che all'epoca suscitarono più interesse erano quelle che trattavano argomenti di politica, spesa pubblica e questioni razziali, con l'affermarsi del genere vi è un crescente elenco delle tematiche trattate. La sua considerazione è aumentata grazie ai diversi premi Pulitzer vinti in questo settore e in particolare a tre inchieste giornalistiche, che ne hanno segnato l'evoluzione. Una di queste riguarda il premio Pulitzer raggiunto nel 1969 da Philip Meyer e il suo team del Detroit Free Press, il metodo scientifico applicato da Meyer ha permesso di dimostrare le reali cause delle rivolte accadute nella 12th strada

di Detroit, da una parte della popolazione di colore. Un'altra inchiesta cardine del giornalismo di precisione è quella di un altro premio Pulitzer, Bill Dedman, nel 1989, che grazie all'ampio uso di tavole, disegni e carte geografiche, molto inusuali in quel periodo, riesce a catturare l'attenzione del pubblico e a diventare un antesignano per tutti coloro che al tempo guardavano ancora con diffidenza i computer. La terza inchiesta è quella di Steve Doig, che nel 1992, in seguito al disastro causato da un uragano porta alla luce le colpe dell'uomo con i suoi abusi edilizi. Doig conduce in modo esemplare l'inchiesta, definita dallo stesso Meyer come l'esempio più importante del giornalismo di precisione.

1.2.1 Philip Meyer "*The people beyond 12th Street*"

"A survey of attitudes of Detroit negroes after the riot of 1967"

Nel 1968 uno staff di giornalisti del Detroit Free Press, diretto da Philip Meyer, si aggiudicò il premio Pulitzer come Giornalismo Locale e di Ultim'ora, realizzando un'inchiesta giornalistica sulla rivolta della 12th strada, avvenuta a Detroit, nel Michigan. Cominciò tutto nelle prime ore del mattino del 23 luglio 1967, con l'irruzione della polizia in un bar, dove 80 uomini e donne di colore stavano celebrando il ritorno di due veterani dal Vietnam. A seguito di questo avvenimento iniziarono una serie di scontri e di disordini pubblici che si trasformano in una delle rivolte civili più letali e distruttive della storia degli Stati Uniti, nell'arco di 5 giorni il bilancio fu di 43 morti, 1.189 feriti, 7.200 arresti e più di 2.000 edifici distrutti, con danni calcolati che si aggirano ai 500 milioni di dollari. Le rivolte hanno una violenza mai vista negli Stati Uniti, i saccheggi si susseguivano e la polizia interveniva con una brutalità devastante, fu necessario anche l'intervento dei militari federali.

Inizialmente gli scontri erano considerati dai media come uno sfogo delle persone più immature, più frustrate, provenienti dal fondo della scala economica che erano prive di altri mezzi di espressione, di origine afroamericana, composta principalmente da persone provenienti dal sud rurale di Detroit. Non tutti però seguivano questa linea di pensiero, la rivolta era troppo violenta per essere considerata come una semplice rivolta

razziale. Lo psicologo e professore dell'istituto di scienze sociali dell'università del Michigan, Nathan Caplan, era convinto che ci fossero delle ragioni ben più profonde riguardo a questa vicenda. Dello stesso parere era Philip Meyer, inviato sul posto dalla *Knight Newspapers*, proprietaria del Detroit Free Press, per documentare la vicenda. Meyer qualche settimana prima aveva concluso, ad Harvard, degli studi sui metodi di analisi dei comportamenti sociali e decise di applicare subito le tecniche apprese. Caplan e Meyer iniziarono quindi a lavorare insieme, decisero di intervistare diverse persone, rivoltosi e non. Dai dati raccolti emerse che gli individui che avevano frequentato l'università avevano le stesse probabilità di partecipare alle rivolte rispetto agli individui che non avevano completato la scuola superiore, smentendo così le teorie sostenute dalla maggior parte dei media, come viene evidenziato nella tabella sottostante.

	Livello d'istruzione		
	Scuola Media	Scuola secondaria	College
Rivoltosi	18%	15%	18%
Non-rivoltosi	82%	85%	82%
Totale	100%	100%	100%

Tabella 1 – Rivoltosità in base al livello d'istruzione

La tabella mostra che il numero dei rivoltosi che hanno frequentato il college è infatti uguale al numero di rivoltosi che si sono fermati alla scuola media. Un'altra teoria, smentita dai dati, affermava che uno dei motivi della rivolta fosse la difficoltà nell'integrazione per le persone di colore cresciute nella zona a sud dalla città, che non erano riuscite ad adattarsi, anche in questo caso il giornalismo di precisione smentì queste teorie scoprendo che la maggior parte dei rivoltosi era cresciuta nella zona nord di Detroit, ben il 73 per cento, mentre solo il 27 per cento proveniva dal sud.

	Dove sei cresciuto da bambino?		
	Sud	Nord	Totale
Rivoltosi	27%	73%	100%
Non-rivoltosi	59%	41%	100%

Tabella 2 – Rivoltosità in base alla provenienza

Questi dati mostrano l'inesistenza di una correlazione tra lo status economico, il livello di istruzione e la provenienza con la propensione alle rivolte. Senza le analisi svolte da Meyer e Caplan probabilmente non sarebbe stato possibile capire le vere ragioni che si celavano dietro ai fatti accaduti a Detroit e che le reali cause delle rivolta erano da imputare principalmente alla brutalità degli interventi della polizia, alle condizioni di sovraffollamento, alla scarsità di alloggi e alla mancanza di posti di lavoro. Con questa inchiesta Meyer dimostra che solo grazie alla raccolta accurata dei dati e al metodo scientifico sia possibile raggiungere una maggior oggettività ed affidabilità nelle notizie.

1.2.2 Bill Dedman "*The Color of Money*"

Il giornalista investigativo e autore americano, Bill Dedman vinse il premio Pulitzer in giornalismo investigativo nel 1989, con *"The Color of Money"* (Il colore dei soldi), una serie di articoli pubblicati sul *The Atlanta Journal and Constitution*, sulla discriminazione razziale da parte di banche e istituti di credito nella città di Atlanta, in Georgia. La prima serie, pubblicata in quattro numeri consecutivi dal primo al quattro maggio del 1988, rivelò che le banche, gli istituti di credito e le casse di risparmio, concedevano prestiti nei quartieri bianchi più poveri della città, ma non facevano altrettanto nei quartieri con una prevalenza di persone di colore, anche se si trattava di persone di classe media o più alta. A parità di reddito, anzi, gli individui bianchi ottenevano un mutuo cinque volte più spesso rispetto agli individui neri, con un divario che aumentava ogni anno. Lo studio dell'*Atlanta Journal-Consitution* di Atlanta analizzò i prestiti bancari per valore di 6.2 miliardi di dollari. Lo studio evidenziò che il fattore che influiva maggiormente per la concessione dei prestiti era la **razza** e non il valore della casa o dei possedimenti. Lo studio mostra che i quartieri

bianchi ricevevano sempre dei prestiti per 1.000 $ per ogni singola famiglia, i quartieri "misti" ricevevano una cifra inferiore e i quartieri neri ricevevano una cifra ancora minore anche includendo i quartieri più benestanti. La ricerca durò cinque anni e furono analizzati un totale di 109.000 prestiti bancari concessi tra il 1981 e il 1986, in 64 quartieri a reddito medio, tra cui 39 quartieri bianchi, 14 neri e 11 misti. Per garantire l'affidabilità dei risultati erano stati scelti dei quartieri che erano comparabili in termini di reddito e di valore dagli alloggi, ovvero tra i 12.849 $ e i 22.393 $.

L'analisi fornì molti spunti di riflessione, prima di tutto che gli uffici più grandi e più importanti erano tutti ubicati in quartieri con una maggioranza di persone bianche e che molte banche o istituti di credito non possedevano nemmeno una filiale nei quartieri neri. Alcune banche avevano deciso di chiudere le proprie filiali nei quartieri che passavano dall'avere una maggioranza bianca ad una maggioranza di persone di colore, altre banche invece rimanevano aperte per più ore nei quartieri bianchi rispetto alle filiali presenti nei quartieri neri. Questa differenza nella concessione dei prestiti aveva interessato alcune delle persone più rilevanti ad Atlanta. Il presidente del *Concilio della città di Atlanta* affermò che questo comportamento era da intendersi come "Razzismo istituzionale", mentre il vice presidente della *Casa federale dei mutui e delle banche di Atlanta*, Robert Warwick, rispose alle domande con indifferenza, affermando che "è ovvio che alcune zone di Atlanta abbiano più difficoltà nell'ottenere il credito, è perfettamente ovvio!"[2].

Nell'immagine sottostante vengono mostrate due mappe: quella a sinistra mostra la distribuzione delle persone bianche e nere ad Atlanta, la parte bianca rappresenta una zona con una maggioranza di persone bianche, la parte nera rappresenta le zone con una maggioranza di persone di colore. Nella figura a destra possiamo vedere le zone dove le banche difficilmente concedono dei prestiti, indicate con il colore nero. Già ad una prima occhiata possiamo notare che c'è una grande corrispondenza tra le zone abitate per lo più da persone di colore e le zone dove difficilmente vengono concessi i mutui.

[2] The Color of Money, http://powerreporting.com/color/.

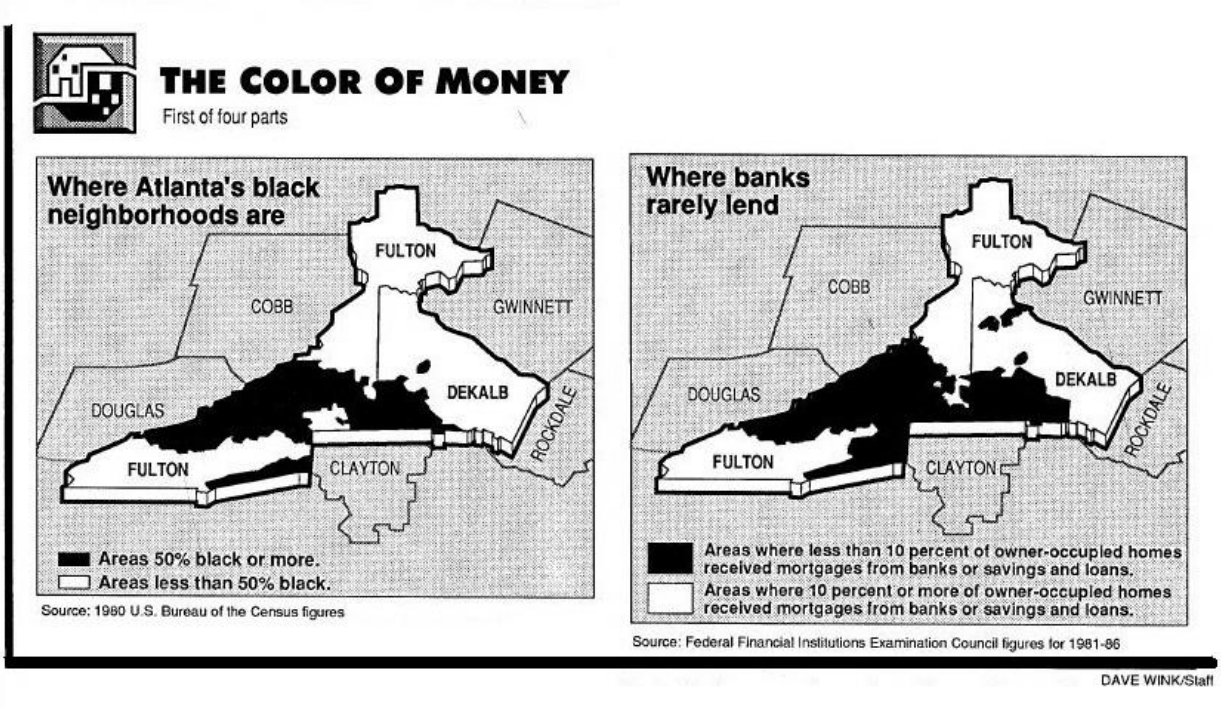


Figura 1 - The Color of Money

Grazie all'uso di mappe, grafici e tabelle che dimostravano in maniera chiara le percentuali e le differenze dei prestiti concessi ai bianchi e ai neri, l'inchiesta suscitò un tale eco da spostare l'attenzione sugli istituti di credito di tutta la nazione. I lettori furono entusiasti del metodo utilizzato che conferiva alla notizia un'oggettività indiscutibile, Dedman fu un pioniere, soprattutto durante quel periodo, dove si guardava ancora ai computer con diffidenza. L'inchiesta non era stata creata solamente con interviste fatte a banchieri, proprietari di case, periti o agenti immobiliari, il giornale aveva ottenuto i rapporti compilati dal *Federal Financial Institution Act* grazie soprattutto al FOIA (*Freedom of Information Act*) e li aveva combinati con i dati demografici del censimento del 1980.

All'inchiesta giornalistica parteciparono alcuni ricercatori universitari appartenenti alla *Johns Hopkins University* e della *Temple University*. Sebbene lo studio non sia stato in grado di includere tutti i mutui concessi o negati, a causa di limitazioni alle leggi di quel periodo e al rifiuto di alcuni istituti di rilasciare per intero i propri dati, la ricerca riuscì a

includere e ad analizzare 82.610 mutui per l'acquisto di una casa e 26.721 prestiti per ristrutturare le propria abitazione, per un totale di circa 6,2 miliardi di dollari in prestiti da parte di banche e istituti di credito.

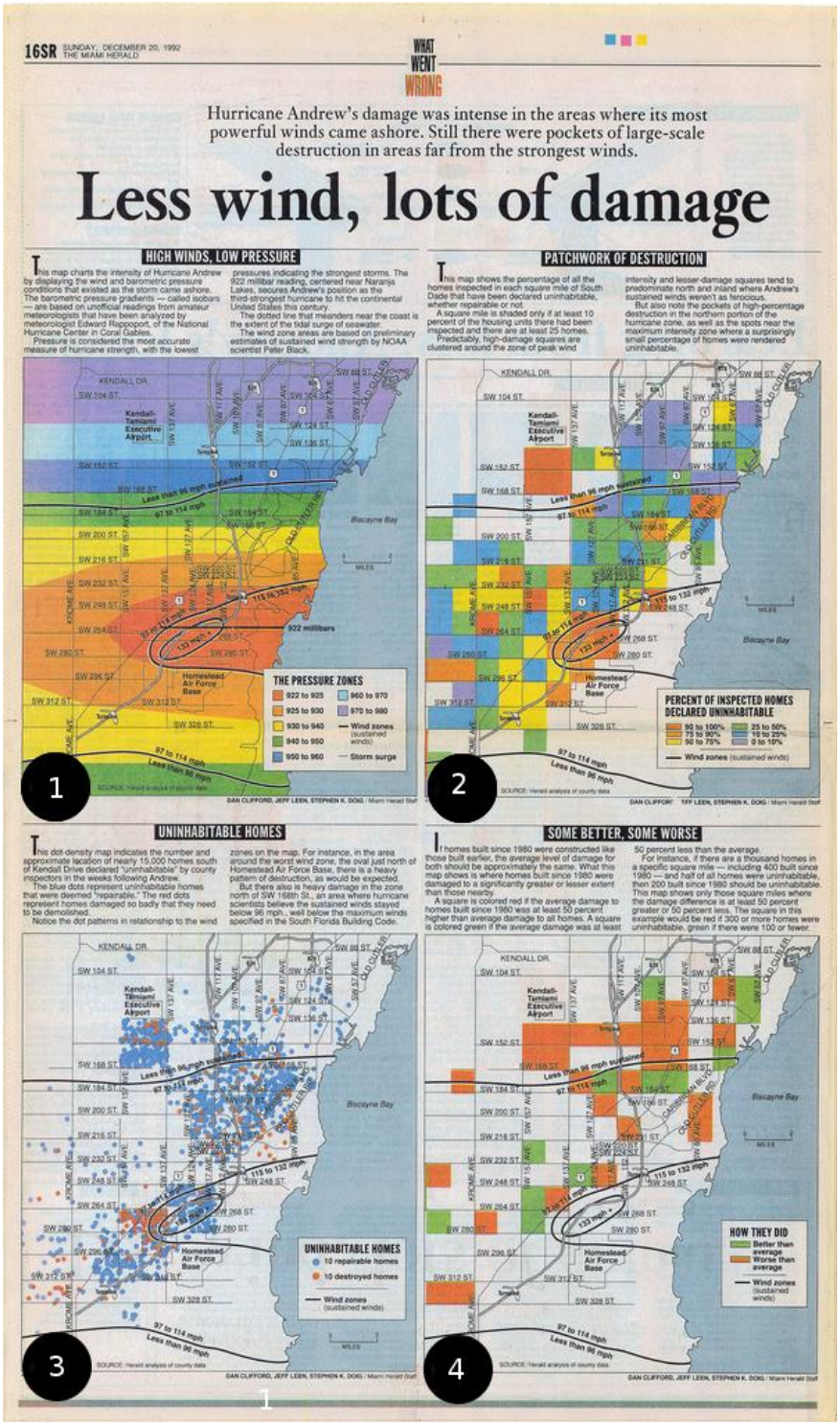
Il lavoro di Dedman rappresenta un passo importante per l'evoluzione tecnologica del giornalismo di precisione. Le sue mappe, il suo modo di rappresentare gli avvenimenti, l'uso del computer per elaborare i dati, erano tutte cose che prima di lui erano usate solo marginalmente, invece dopo il suo arrivo molti giornalisti hanno deciso di seguirlo e di fare un uso sempre più ampio del computer e degli strumenti digitali.

1.2.3 Stephen K. Doig "*What Went Wrong*"

Nel 1992 Stephen K. Doig professore presso la scuola di giornalismo dell'università dell'Arizona, rimase coinvolto nella tragedia provocata dall'uragano Andrew. L'uragano venne classificato come il secondo più distruttivo nella storia degli Stati Uniti e fu l'ultimo uragano di categoria 5 che colpì gli Stati Uniti durante il XX secolo. Provocò la morte di 65 persone e danni per 26.5 miliardi di dollari, la maggior parte dei quali vennero registrati nella Florida meridionale, in particolare nella città di Miami dove, tra l'altro, venne devastata anche la casa di Doig. Dalle fotografie aeree del disastro emerse un quadro chiaramente disomogeneo, nel quale fu possibile notare che le costruzioni avevano subito dei danni che non dipendevano dalla direzione e dalla forza del vento, ma dalla suddivisione territoriale. A seguito dell'accaduto Doig, con una squadra di giornalisti del Miami Herald, decise di realizzare l'inchiesta *"What Went Wrong"*, uno speciale report di 16 pagine nel quale vennero analizzare e confrontate le registrazioni metereologiche, le ispezioni degli edifici e i danni riportati dagli edifici, per dimostrare che la principale causa dei danni era da attribuirsi agli abusi e alle frodi edilizie, che avevano portato a realizzare nel tempo case sempre meno sicure, per risparmiare sui materiali di costruzione.

Dalle riprese aeree, emerse che i danni seguivano una linea che non dipendeva dalla forza del vento e che la bassa qualità delle costruzioni e dei materiali usati avesse trasformato una devastante tempesta in uno dei più grandi disastri nella storia degli Stati Uniti. Durante 4 mesi di investigazione la Miami Herald analizzò i danni riportati da più di 60.000 case,

per arrivare a 2 grandi scoperte: La prima era che molte delle case cadute presentavano dei difetti nascosti, una volta registrati i danni si scoprì che la parte della città più danneggiata non coincideva con la parte dove si erano abbattuti i venti più forti. La seconda grande scoperta fu che le case più vecchie se la l'erano cavata molto meglio. Venne stimato che chi possedeva una casa costruita a partire dal 1980 aveva il 68% di probabilità in più di rimanere senza casa rispetto a chi viveva in case più vecchie. Grazie ad alcune rappresentazioni grafiche, l'inchiesta di Steve Doig riuscì a chiarire il ruolo e le responsabilità dell'uomo nel disastro. Dalla mappa sottostante risulta molto facile, per i lettori, capire che molti dei danni causati agli edifici sono da imputare alla cattiva qualità dei materiali usati, e non al vento. In particolare viene dato molto risalto alla scadenza degli edifici costruiti dopo il 1980, rispetto a quelli più vecchi, in altre parole l'essere umano era responsabile di una parte considerevole del disastro.



Nella figura a sinistra il riquadro n°1 indica la forza distruttiva del vento. L'immagine n°2 in alto a destra mostra invece la percentuale di case dichiarate inagibili. Il riquadro n°3, in basso a sinistra mostra con i puntini blu le case riparabili e con i puntini rossi le case non riparabili. L'ultima immagine, la n° 4, che forse è quella più significativa, mostra come hanno resistito all'uragano le case costruite dopo il 1980.

Figura 2 - Mappa dei venti e dei danni

1 – Il rosso indica le zone con i venti più forti.
2 – Il colore rosso indica una più alta percentuale di case inagibili
3 – I puntini rossi indicano le case non riparabili, quelli blu le case riparabili
4 – Le zone arancioni indicano le zone dove le case moderne hanno subito più danni rispetto alle vecchie.

Successivamente un'altra indagine metteva a confronto la percentuale di case ispezionate, la percentuale di case ritenute inagibili, l'anno medio di costruzione e il valore medio valutato, arrivando alla conclusione che nelle aree con i venti più lievi, da 85 a 127 miglia all'ora, le case più nuove (costruite dopo il 1979) avevano una probabilità tre volte maggiori di diventare inagibili rispetto a quelle costruite in precedenza. In questo modo, grazie ai dati raccolti e all'aiuto del computer, Doig dimostrò come la corruzione dell'edilizia urbana aveva modificato a proprio favore i regolamenti edilizi, costruendo negli anni case sempre meno sicure, ad alto rischio di crollo con l'arrivo di un disastro naturale. Dopo l'evento dell'uragano, Doig non si limitò a fare delle semplici constatazioni ma andò oltre, procurandosi tutti i dati per dimostrare la sua tesi, usando appieno il metodo scientifico.
Nel libro del 1993, *The New Precision Journalism*, Philip Meyer ha indicato questo articolo come l'esempio più importante e più rappresentativo del giornalismo di precisione. Tanto che il modo di procedere di Steve Doig è diventato un esempio da seguire, un esempio del tipo di servizio pubblico che dovrebbe offrire il giornalismo.

2 IL DATA JOURNALISM OGGI

Dalla nascita del World Wide Web, nel 1991, la tecnologia digitale ha invaso tutti i campi, creando dei grandi cambiamenti in tutti i settori: scuola, lavoro, università, tempo libero e tanti altri. Questa "rivoluzione" ha cambiato anche il data journalism, chiamato in questo modo proprio per sottolineare l'uso massivo dei dati. Negli esempi precedenti abbiamo visto come Meyer, Dedman e Doig, siano scesi sul campo a raccogliere dati e informazioni per scovare la verità, che altrimenti sarebbe rimasta nascosta. Oggi il modo di lavorare è diverso, è molto raro che chi si occupa di data journalism vada direttamente sul campo a raccogliere i dati. Anche se ci sarà sempre bisogno di qualcuno che raccolga i dati sul campo, grazie ad internet, abbiamo un portale universale dove chiunque può accedere alle informazioni e chiunque vi può contribuire. Oggigiorno siamo sommersi dalle informazioni, la maggior parte delle informazioni che un data journalist necessita per svolgere le sue inchieste esistono già, il problema è solo riuscire a trovarle su internet. In questa parte vediamo quali sono i giornali che sono più attenti al data journalism, come sono cambiati i metodi di lavoro dei giornalisti e quali sono gli strumenti principali.

2.1 LE REDAZIONI PIÙ ALL'AVANGUARDIA

Negli ultimi anni molte redazioni giornalistiche hanno deciso di dare sempre più spazio al data journalism. Tra le grandi redazioni che recentemente hanno contribuito allo sviluppo del data journalism troviamo le inglesi *The Guardian* e *BBC* (British Broadcasting Corporation), l'australiana *ABC* (Australian Broadcasting Corporation), la società svizzera *Zeit Online*, le americane *New York Times*, *Washington Post*, *ProPublica* e *FiveThirtyEight* e l'argentina *La Nación*. Tra le realtà italiane possiamo citare *Dataninjia*, *Wired Italy* e *Limpido* che appartiene al gruppo editoriale *L'espresso.*

2.1.1 Il datablog del The Guardian

Il The Guardian può essere considerato come la testata giornalistica più innovativa per quanto riguarda il data journalism, è stata la prima a pubblicare e rendere disponibili al

pubblico i dati delle inchieste giornalistiche effettuate. Il suo "datablog" è un punto di riferimento per tutte le più importanti organizzazioni che si occupano di news. A detta di uno dei più autorevoli giornalisti del The Guardian, *Simon Rogers*, il datablog inizialmente doveva essere solamente un piccolo blog nel quale condividere i dati appartenenti alle notizie che venivano pubblicate sul giornale, ora invece contiene una grande quantità di dati, nel quale tutte le persone che hanno un accesso a internet possono controllare dati di ogni genere. Uno degli eventi più importanti per lo sviluppo del data journalism è avvenuto nel 2009, quando il presidente degli Stati Uniti, Barak Obama, ha aperto gli archivi dei dati del governo, in particolare i dati che riguardano il budget e la spesa del governo. Il suo esempio è stato presto seguito da tutto il mondo, poco tempo dopo anche Australia, Nuova Zelanda e Gran Bretagna hanno pubblicato i dati della propria spesa pubblica. Il Guardian fu il primo giornale ad usare questi dati per fare delle indagini giornalistiche, grazie a queste pubblicazioni, tra le altre cose, è riuscito individuare e pubblicare lo scandalo delle spese pazze dei parlamentari Britannici, dando la possibilità a tutti gli utenti di poter sfogliare i dati relativi alle spese che erano presenti nel datablog. Ma l'evento che ha cambiato il modo di lavorare al The Guardian e ha permesso la diffusione a livello mondiale del data journalism è arrivato nella primavera del 2010, con lo scandalo dei registri di guerra di "WikiLeaks". Nel quale vennero pubblicati dei documenti dettagliati che riguardavano i diversi fallimenti dei militari americani in Afghanistan, seguiti poco tempo dopo dai dati relativi alla guerra in Iraq. Ancora Rogers, spiega come l'organizzazione delle notizie dipenda molto dalla distanza con il "tavolo delle news" e come lo scandalo WikiLeaks ha influito fortemente anche con la loro organizzazione interna. Rogers dice che prima di WikiLeaks i collaboratori del datablog stavano in un altro piano rispetto agli altri giornalisti e al tavolo delle news, avendo un potere decisionale quasi nullo, mentre dopo lo scandalo si sono sistemati nello stesso piano dei giornalisti e ora si trovano vicino al tavolo delle news, questo significa che per loro è molto più facile suggerire nuove idee ai reporter ed aiutarli con le storie. Il datablog del The Guardian è ora uno dei più grandi e completi, anche se inizialmente c'erano delle persone che nutrivano dei dubbi, in tanti si chiedevano perché la gente avrebbe voluto analizzare i dati senza nessuna spiegazione. Ancora Rogers spiega come da

quando il datablog è nato il ruolo dei giornalisti all'interno del The Guardian stia cambiando, stanno diventando una sorta di interpreti che aiutano la gente a capire i dati ed anche a pubblicarli.

2.1.2 ProPublica

ProPublica è una corporazione no profit nata nel 2007, dall'idea dei coniugi *Herbert* e *Marion Sandler,* entrambi amministratori delegati della società *Golden West Financial*. Dopo aver messo sotto contratto *Paul Steiger*, ex caporedattore del *Wall Street Journal*, gli assegnano il compito di creare e organizzare l'ossatura di ProPublica, che, in pochi anni, diventa una delle più importanti aziende di data journalism. Si definisce come una società di giornalismo investigativo e, a differenza dei blog citati precedentemente, questa nuova corporation non è di proprietà di un giornale e pubblica i suoi articoli solamente online, in quanto non possiede una licenza commerciale. La società viene finanziata principalmente dalla Golden West Financial, con 10 milioni di dollari annui, ma anche da altre società internazionali, come la *Knight Foundation,* la *MacArthur Foundation* e la *Ford Foundation*. Nella sua breve vita ProPublica ha già collaborato con molti dei giornali più importanti, dato che non possiede una edizione cartacea i suoi articoli vengono spesso pubblicati, senza nessun costo per le redazioni, su molti importanti giornali, come il New York Times, *Usa Today, NewsWeek, Los Angeles Times* e le notizie vengono mostrate su alcuni canali televisivi, come la *CNN*. Ha ricevuto diversi premi per le sue inchieste giornalistiche, nel 2010, con la collaborazione del New York Times, riceve il premio Pulitzer come miglior giornalismo investigativo, nel 2011 vince il suo secondo premio Pulitzer, i reporter *Jesse Eisinger e Jake Bernstein*, vincono il premio di *Giornalismo Nazionale*, per la loro serie di notizie, *The Wall Street Money Machine* (Wall Street, la macchina da soldi). Inoltre ha ricevuto molte lodi per l'inchiesta sulla compagnia *Psychiatric Solutions,* condotta con il *Los Angeles Times*, nella quale si scoprì che un'azienda ospedaliera del Tennessee accumulava gran parte del suo profitto comprando ospedali fallimentari e licenziandone i membri. Nonostante nella considerazione generale ProPublica viene vista come un'organizzazione con un alto livello di oggettività, anche lei ha ricevuto delle pesanti critiche da questo

punto di vista. Una particolarmente pesante fu quella di *Dave Kopel,* analista per l'istituto Cato, ex editorialista per la *Rocky Mountains News*. Kopel in un report critica i dati sulla fratturazione idraulica mostrati da Propublica, formulando l'accusa di mostrare una "*serie unilaterale di fatti disposti a sostenere un predeterminato punto di vista*". Kopel sosteneva infatti che ProPublica tendeva a sviluppare costantemente degli argomenti per dimostrare che *"Il governo non sta facendo una lavoro abbastanza buono per controllare gli affari, soprattutto quelli che riguardano nuove aziende"*. Ad oggi ProPublica rappresenta uno dei migliori esempi di data journalism e la sua redazione, che si occupa solo di investigazioni di questo genere, riesce a produrre sempre degli articoli di grande qualità ed obbiettività, rivendo anche quest'anno un riconoscimento dalla *Global Editors Networks.*

2.1.3 FiveThirtyEight

FiveThirtyEight è nato nel marzo del 2008, fu lanciato da Nate Silver. Dal giugno del 2013 pubblica le notizie per conto dell'emittente televisiva americana ESPN. La maggior parte della sua notorietà è arrivata nel periodo in cui pubblicava le notizie per conto del New York Times, tra l'agosto del 2010 e il Giugno del 2013. Questo blog ha ricevuto degli importanti riconoscimenti: due *Bloggie Awards* uno nel 2008 per la miglior copertura politica e uno nel 2009 per il migliore blog sulla politica, due *Webby Awards* come "miglior blog politico", nel 2012 e nel 2013. La crescita della fama di FiveThirtyEight è legata a due eventi in particolare: le elezioni politiche del 2008 e del 2012. Nel 2008 Nate Silver e i suoi metodi di predizione riuscirono a prevedere il risultato delle elezioni politiche americane, in 50 stati su 51, sbagliando la previsione solo per lo stato dell'Indiana. Per questa previsione Silver usò un metodo di predizione unico, derivato dalla sua esperienza nella "sabermetrica", ovvero l'analisi del baseball attraverso le statistiche, un metodo che, a suo dire, "*bilancia i risultati dei sondaggi con i dati demografici*" nel quale lui pesa ogni sondaggio, basandosi sull'andamento storico dei sondaggi, sulla dimensione del campione, e sull'ordine temporale dei sondaggi. Questa ottima previsione aumentò la fama di Nate Silver come predittore, ma il vero capolavoro arrivò nel 2012, in occasione delle elezioni politiche, nel quale, lui e il suo staff, furono in grado di predire il risultato

delle elezioni politiche in 50 stati su 50 e nel District of Columbia, nel periodo delle elezioni il blog FiveThirtyEight riuscì a raggiungere un numero di visite pari al 20% del totale del New York Times[3].

La qualità delle previsioni di FiveThirtyEight emergono non solo nella politica ma anche in altri settori, nel 2013 riuscì a prevedere che i *"San Francisco 49ers"* avrebbero vinto il *Super Bowl*, sempre nello stesso anno riuscì a prevedere ben tre vincitori su quattro degli *Academy Awards*. Nate Silver e il suo staff, soprattutto grazie a queste grandi previsioni, dimostrano di essere tra i migliori nel campo dell'analisi dei dati, e dalla loro esperienza nella previsione, nel 2012 è uscito il libro *The Signal and the Noise* (Il segnale e il rumore), nel quale vengono coperti degli argomenti riguardanti molte aree di previsione, incluso scommesse sportive, previsioni scientifiche, previsioni del tempo, economia e poker, nel quale è richiesta l'abilità di prevedere il comportamento dell'avversario. Una parte molto interessante del suo libro è quella parla del programma chiamato "PECOTA", il quale usa dei metodi di analisi per prevedere i risultati nel baseball.

Questo blog è stato spesso al centro delle discussioni riguardanti il data journalism: da una parte, con una visione positiva, per la sua competenza e per gli ottimi risultati ottenuti dalle sue previsioni. D'altra parte, molti data journal, hanno criticato il comportamento di FiveThirtyEight per il fatto di non pubblicare i dati che vengono usati per la creazione delle notizie.

2.1.4 "The Upshot" - New York Times

Il sito web del New York Times riceve circa 31 milioni di visite ogni mese, l'edizione cartacea produce circa un milione di copie al giorno, questo lo rende uno dei giornali più apprezzati al mondo. Dal 22 aprile del 2014, il sito web del NY Times, si è arricchito di una nuova sezione dedicata al data journalism, *The Upshot*, che va a prendere il posto del blog di Nate Silver, FiveThirtyEight, il quale per quasi tre anni ha pubblicato le notizie per conto nel New York Times.

[3] Fonte: http://qz.com/185922/the-upshot-is-the-new-york-times-replacement-for-nate-silvers-fivethirtyeight/

The Upshot, a detta del suo caporedattore *David Leonhardt*, ha l'obbiettivo di facilitare la comprensione delle storie che vengono pubblicate sul NY Times attraverso l'analisi dei dati, principalmente perché spesso i classici articoli non permettono ai lettori di comprendere a pieno tutti i dati e la gente non riesce a capire le notizie come vorrebbe. In quest'ottica il data journalism permette non solo di informare meglio, ma permette ad ognuno di avere un visione delle notizie più obbiettiva. The Upshot mira ad agevolare la comprensione delle notizie soprattutto tramite una *data visualization* chiara e facilmente leggibile, cercando di creare una sorta di collaborazione tra i giornalisti e i lettori, in modo che i secondi possano notare quello che i giornalisti non hanno ritenuto importante.

2.2 GLI STRUMENTI DEL DATA JOURNALIST

Per comprendere appieno il data journalism c'è la necessità di soffermarsi sugli strumenti più usati in questo settore. Naturalmente non tutti utilizzano gli stessi strumenti ogni giornalista ha le sue preferenze. Proprio per migliorare la qualità degli strumenti usati dai giornalisti *The Editor Lab* organizza periodicamente dei meeting tra giornalisti e programmatori, chiamati *hackatones*. Al momento i data journalist usano una grande quantità di strumenti, come quelli impiegati per raccogliere i dati dalla rete, attraverso quello che viene definito scraping, tool per la conversione di file da immagini o pdf in formati facilmente elaborabili, come CSV, le API disponibili gratuitamente in rete, che permettono a diversi software di comunicare tra di loro. Svolgono un'importante funzione anche i software gratuiti che permettono l'elaborazione, l'analisi statistica e strumenti per la visualizzazione dei dati. In questa parte possiamo analizzare quelle che sono le azioni che vengono svolte più spesso dagli esperti del mondo del data journalism.

2.2.1 Gli Spreadsheet (fogli elettronici)

In questo momento i fogli elettronici sono un po' antiquati, pian piano il loro uso da parte dei data journalist sta diminuendo a favore di altri strumenti. Spesso però costituiscono una buona basa di partenza per iniziare il lavoro, in virtù della loro versatilità, infatti, sono in grado di operare su diversi formati, i data journalist sono soliti salvare i dati con il

formato *plain-text*, cioè come un file di testo con valori determinati da virgole, come CSV. Un altro elemento che ha favorito la diffusione dei fogli elettronici è dato dal fatto che esistono una grande varietà di programmi che permettono di elaborare questi fogli elettronici, sia commerciali, come il famosissimo *Microsoft excel* che è forse quello più completo, o è possibile scaricarne qualcuno con una licenza free, che contiene quasi le stesse proprietà, come il programma *calc*, della suite *libreoffice*. Negli ultimi tempi c'è stato un grande sviluppo dei programmi *web-based*, ovvero dei programmi che possono essere eseguiti direttamente su internet attraverso il proprio browser, come la suite gratuita di google drive, molto apprezzata anche dai professionisti, è gratis e permette il lavoro collaborativo.

2.2.2 DBMS (Data Base Management System)

Uno degli strumenti più usati in assoluto sono i database, più precisamente i *DBMS* (Data Base Management System). Infatti, dopo che si lavora con i fogli elettronici per un po' di tempo, emergono delle limitazioni e molti data journalist decidono di sfruttare le potenzialità dei sistemi di gestione dei database, soprattutto quando c'è bisogno di incrociare i dati di più fogli elettronici e ci sono molti dati da interrogare. Quasi tutti i DBMS relazionali usano il linguaggio *SQL* (Structured Query Language), che permette di descrivere con esattezza il sottoinsieme di dati che si vuole estrarre e i precisi cambiamenti da apportare, inoltre permette di effettuare queste "interrogazioni" attraverso gruppi di dati correlati. *MySql* in questo campo è forse il DBMS gratuito più usato al mondo. Oltre a questo troviamo *Access*, che fa parte della suite office e quindi è disponibile solo a pagamento, in maniera minore vengo usati altri programmi come *PostreSQL, SQLite,* o *Base* che fa parte della suite LibreOffice. Questi programmi rispetto ai fogli elettronici hanno il vantaggio di poter lavorare con diversi set di dati contemporaneamente e permettono al sistema di rimanere performante anche quando la mole di dati su cui si lavora è molto grande.

2.2.3 Strumenti per la pulizia dei dati

Molto spesso i dati non si presentano nel modo in cui noi vogliamo, quindi per ripulire i dati e trasformarli in un formato utile ci potrebbero servire diversi strumenti, a

seconda di quello di cui abbiamo bisogno. Infatti può capitare di dover assemblare due tipi di dati che sono formattati in modo diverso, se prendiamo ad esempio il formato delle date, in Italia e nel resto d'Europa (eccetto il Regno Unito) vengono scritte con la formula giorno/mese/anno mentre negli Stati Uniti, Canada e Regno Unito il formato più diffuso è mese/giorno/anno. Un'altra differenza, stavolta nella punteggiatura, potrebbe esserci quando si esaminano dei numeri decimali, in Italia i numeri decimali vengono generalmente scritti usando la virgola (es. 3,14), mentre lo stesso numero negli Stati Uniti verrebbe scritto con il punto (3.14). Un altro problema potrebbe essere quello di avere del testo in una colonna che dovrebbe contenere solamente una data, ad esempio si potrebbe avere "fine 2013" invece del solo numero 2013. Tutti questi sono dei casi dove dobbiamo intervenire per modificare i dati. Quando i dati da modificare sono di piccola mole si può intervenire anche a mano modificando tutte le tabelle, quando invece dobbiamo pulire migliaia o anche centinaia di righe (o record), è opportuno affidarsi a degli strumenti che permettono di facilitare il lavoro. Uno dei più apprezzati è *Google Refine*, che si presenta esteticamente come un foglio elettronico e permette di operare sofisticate trasformazioni di dati. Molto apprezzato anche *Data Wrangler*, che presenta molte funzionalità del programma precedente. Oltre a questi programmi web-based esistono inoltre delle ottime soluzioni fornite dal proprio sistema operativo *grep, find* o *sed*, se ci troviamo in ambiente Unix. Anche la suite *CSVKit* offre un set di strumenti straordinari, è stato sviluppato da un team di giornalisti ed è molto utile per imparare a lavorare su linea di comando.

2.2.4 Strumenti per la data visualization

La data visualization nel data journalism assume un ruolo fondamentale, una buona visualizzazione dei dati è lo scopo finale di ogni progetto di data journalism. Uno dei primi obbiettivi della data visualization è quello di comunicare in modo chiaro ed efficiente le informazioni. Attraverso le infografiche la comprensione delle informazioni viene resa più facile e immediata, facilitando la comprensione dei dati da parte dell'utente. La maggior parte degli strumenti per leggere i fogli elettronici dispongono di diagrammi e grafici. Tra i programmi presenti sul web i più diffusi sono *Google Fusion Table* e *Tableu Public*, sono

entrambi software che non richiedono la conoscenza di un linguaggio di programmazione, facili da usare, anche se il primo (Fusion Table) ha bisogno di più manualità, entrambi offrono dei risultati sorprendenti, vengono usati spesso anche dai professionisti, infatti, i giornalisti del The Gardian usano spesso Google fusion table. Oltre a questi due, negli ultimi tempi sta salendo alla ribalta datawrapper, che si propone come uno strumento multifunzionale capace di essere utile in diversi settori del data journalism. Questi sono degli strumenti semplici, intuitivi e il loro funzionamento è facile da apprendere, ma in alcuni casi c'è la necessità di usare un programma con più funzionalità, più flessibile e potente. In questi casi è necessaria la conoscenza di alcuni linguaggi di programmazione, molto dei più recenti software sono basati sul linguaggio Javascript. *The R project* è un tool molto potente che combina strumenti di visualizzazione e analisi dati, ha un linguaggio "proprio" basato sul linguaggio C/C++. Anche *highcards* si è rivelato uno strumento potente, versatile e non troppo complicato da usare, anche se sono necessarie basi del linguaggio Javascript per poterlo usare. Tra gli altri strumenti, sta riscuotendo molto successo, la libreria di javascript *D3.JS*.

2.2.5 Mappe interattive

Una funzione molto ricercata dai giornalisti è quella di ricreare delle mappe con le quali gli utenti possano interagire. Usando questi programmi è possibile visualizzare i dati su una fedele mappa. Tra i programmi più diffusi troviamo *CartoDB*, un programma molto intuitivo che consente di creare mappe molto funzionali, la visualizzazione che si crea con questo strumento è davvero molto piacevole, è disponibile solo a pagamento. *Google maps*, attraverso l'uso di alcune api, permette di personalizzare la visione delle proprie mappe in modo da mostrare i dati. *JVectorMap* e *Leaflet* sono due strumenti molto potenti, hanno il vantaggio di essere gratuiti, ma necessitano di un po' di manualità e conoscenza di JavaScript. Tra gli altri strumenti troviamo *Geofeedia* che permette di visualizzare una mappa degli stati o degli *hashtag* postati dagli utenti sui social network.

Un'altra funzionalità molto apprezzato dagli utenti è quello di mostrare i dati attraverso una rappresentazione geospaziale, questo permette agli utenti di navigare con molta più naturalità tra i dati.

2.2.6 Linguaggi di scripting

I dati spesso non vengono pubblicati nella maniera corretta o nei formati più congeniali e in questi casi c'è la necessità di ricorrere al fai da te, bisogna costruirsi da soli il proprio tool. C'è una grande differenza di potenziale tra l'essere un semplice utente di software o un progettista, per questo gli attuali data journalist sono quasi obbligati a conoscere almeno un linguaggio di scripting. Tra i linguaggi di scripting c'è una grande varietà di scelta, in questo momento *Python* e *Ruby* sembrano essere i preferiti tra i giornalisti, anche se *PHP* e *Perl* rimangono i due linguaggi di scripting più usati, soprattutto dai professionisti. Conoscere uno di questi programmi di scripting permette di creare degli strumenti in grado di effettuare il *web scraping*, un'operazione quasi fondamentale per il data journalist, che permette di estrarre i dati da pagine web non strutturate. Questa è una funzionalità molto importante per i data journalist dato che molto spesso i dati delle pubbliche amministrazioni non vengono rilasciati in formati elaborabili. Spesso infatti i dati pubblicati su una pagina web non sono scaricabili o ne viene permesso il download solo in formati scomodi, come il PDF, definito da alcuni esponenti del The Guardian come il peggior formato dati conosciuto e da altri esperti come il peggior nemico del data journalism. Naturalmente anche per il web scraping ci sono dei tool web-based che ci possono facilitare la vita, il primo fra tutti è *Scraper*, che è una estensione per il conosciuto browser Chrome, questa estensione ci permette per esempio di estrarre una tabella, da pagine web non strutturate con un semplice click e di esportare questi dati in formati attraverso la suite Google Docs. Oltre a questo possiamo citare programmi come *Outwit Hub*, una estensione per firefox, o *Scraperwiki* molto utile per chi non possiede nessuna conoscenza dei linguaggi di programmazione.

2.2.7 Strumenti di analisi dei documenti

Uno dei principali obbiettivi del giornalismo attuale è la trattazione come dati di documenti di grandi dimensioni, in questo settore troviamo programmi come *DocumentCloud* che fornisce una pratica interfaccia per esaminare documenti in formato PDF, permettendo la ricerca nel documento e l'estrazione dei punti d'interesse. Un altro programma molto utile è *Jigsaw*, che permette di navigare attraverso una notevole mole di documenti. Il giornalismo non si è ancora spinto tantissimo in questo campo, quindi è possibile che nel prossimo periodo escano dei programmi veramente interessanti e rivoluzionari.

2.2.8 Data Warehousing

Il data warehouse è uno archivio informatico utile per unire dati appartenenti a diverse risorse informatiche. In genere vengono usati dalle aziende per facilitare la fase di analisi dei dati. Questi archivi uniscono i dati presenti su diversi database basandosi su una "chiave", cioè un campo comune ai diversi database. Unificare i dati provenienti da diversi database non è un lavoro semplice, specialmente quando ci sono più persone di una stessa organizzazione che stanno effettuando delle analisi contemporaneamente.
I data warehouse rappresentano uno strumento molto utile ai data journalist, permettendogli di svolgere con relativa semplicità delle operazioni di analisi dei dati che altrimenti sarebbero molto più complesse.

2.2.9 Big Data

Il termine *Big Data* si riferisce a qualsiasi set di dati cosi grande e complesso che è difficilmente trattabile attraverso procedure tradizionali, in questo caso vengono richiesti degli strumenti non convenzionali per estrapolare, gestire e processare le informazioni in un tempo ragionevole. Non esiste una dimensione di riferimento, che identifichi un set di dati come Big Data, questo perché i computer sono sempre più veloci e i dati sono sempre più grandi. Con i big data la mole di dati trattati e dell'ordine degli *Zettabyte* (migliardi di Terabyte), per gestire questa mole di dati, i normali *database management system* non sono sufficienti, per questo sono necessari decine, centinaia o addirittura migliaia di server che

lavorino in parallelo. Il concetto di Big Data è molto relativo, quello che un'azienda può considerare Big Data può non avere la stessa considerazione per un'altra azienda. Il data journalism ha uno stretto rapporto con i big data, uno dei compiti principali dei giornalisti è spesso analizzare questi big data e dargli un senso. Tra le principali applicazioni open source usate per la gestione dei big data tra le più diffuse troviamo *NoSql* e *Apache Hadoop*, usata anche dai due giganti del web, Facebook e Yahoo, un'altra applicazione molto importante è *MapReduce* creata dalla google. Queste applicazioni effettuano l'analisi dei Big-data suddividendoli in parti e mandandoli in esecuzione in parallelo su diversi server.

3 DJA - DATA JOURNALISM AWARDS

Negli ultimi anni il data journalism ha raggiunto una tale importanza che molte testate giornalistiche hanno deciso di orientarsi verso questa filosofia e questo ha spinto i rappresentanti delle più importanti testate giornalistiche mondiali a istituire un'associazione per premiare le migliori inchieste di data journalism. La *GEN (Global Editors Network)*, è una comunità che conta più di 1000 caporedattori che rappresentano più di 80 stati e 300 gruppi mediatici in tutto il mondo, il cui obbiettivo è quello di incentivare l'uso e lo sviluppo delle nuove tecnologie digitali da parte delle organizzazioni giornalistiche, favorendo lo sviluppo di un giornalismo di alta qualità. Tra le iniziative più importanti troviamo il "*The Editors Lab*" e il "*DJA*" (*Data Journalism Awards*). Il data journalism Awards è il primo contest internazionale che, con dei riconoscimenti in denaro, premia i migliori lavori nel campo del data journalism. La sua storia è molto recente, è stato istituito nel 2012 avendo sin da subito un enorme successo. Questi premi sono finalizzati ad incoraggiare la nascita di nuovi standard nel campo del data journalism. Il The Editors Lab è un programma basato su una serie di meeting occasionali, generalmente chiamati "hackaton" o "hack day". Tra gli obbiettivi principali c'è anche quello di creare un dialogo tra ingegneri, informatici e chi lavora nel mondo del giornalismo, soprattutto per sviluppare nuove tecnologie che possano essere utili a chi lavora nel data journalism. Il GEN, oltre ad essere finanziato dalle più importanti redazioni giornalistiche mondiali, è sponsorizzato da due colossi del mondo di internet, Google e Yahoo. Oltre a questi due concorsi nell'aprile del 2014 sono iniziati i corsi sul data journalism, i corsi vengono svolti in lingua spagnola e francese oltre che in inglese.

3.1 LE MIGLIORI INCHIESTE GIORNALISTICHE DEL 2014

Nel 2014 i Data Journalism Award hanno premiato un totale di otto categorie, ognuna di queste ha ricevuto un compenso di 2000 euro. Tra i vincitori figura anche un progetto italiano, mentre tra i 75 finalisti erano presenti ben tre inchieste giornalistiche

appartenenti a redazioni italiane. Le categorie premiate in questa sfida sono state otto, nella giuria hanno trovato spazio dei giornalisti e redattori di caratura internazionale, tra i quali possiamo individuare il direttore dalla comunicazione e degli affari pubblici di Google, il presidente di ProPublica e tante altre persone molto rilevanti nel settore giornalistico.
Ecco l'elenco degli otto premi assegnati.

1. **Best Story on a single Topic** (Miglior storia per un singolo tema): Detective.io - *The Migrants' Files.*

   Progetto ideato in Italia da *dataninjia.it* e che ha coinvolto diverse redazioni europee, come Journalism++, SAS e Journalism Stockholm. È stato sviluppato grazie alla piattaforma *detective.io*. L'articolo è stato pubblicato in Italia con il titolo *Mar Mediterraneo, tomba di migranti*.
2. **Best Data-driven Investigation** (migliore inchiesta giornalistica basata sui dati) – The Washington Post, con *Homes for The Taking: Liens, Loss and Profiteer.*

   Nel 2013, il Washington Post ha scoperto un antico sistema Americano di recupero delle tasse che consente agli investitori di prendere centinaia di case a prezzo stracciato per poi rivalutarle ad un prezzo molto più alto. In questo modo tantissime persone si sono ritrovate senza una casa, pieni di debiti e nella disperazione.
3. **Best Data Visualization** (migliore nella visualizzazione dei dati) – The New York Times con *Reshaping New York* (rimodellazione di New York).

   Permette di effettuare un viaggio interattivo attraverso la storia dello sviluppo della grande Mela, dai grattacieli alle piste ciclabili che hanno cambiato il volto della città.
4. **Best Application or Website** (miglior sito o applicazione web) – La Nación con il sito web *declaraciones juradas abiertas* (dichiarazioni giuridiche aperte).

   I giornalisti argentini de "La Nación" hanno lottato per contrastare la mancanza di un FOIA in Argentina e hanno creato un'applicazione web che permette ai cittadini di conoscere quanto i loro politici si sono arricchiti negli anni. Con oltre 30 volontari che hanno lavorato per inserire i dati, La Nación è stato il primo grande lavoro di open data nel Sud America.
5. **Best Individual Portfolio** (miglior raccolta dati individuale) – *Chad Skelton*.

Il giornalista del quotidiano canadese *The Vancouver Sun* ha presentato il suo lavoro personale, con diversi articoli riguardanti le notizie di attualità in Canada, come un foglio di calcolo per capire quanto possono risparmiare le persone comuni, oppure visualizzare le abitudini dei cittadini nell'uso dei trasporti pubblici. Skelton è riuscito secondo la giuria a creare dei lavori di alta qualità.

6. **Best Team or Newsroom Portfolio** (miglior raccolta dati per un gruppo o una redazione) – NZZ (acronimo di *Neue Zürcher Zeitung*), società svizzera che si occupa di raccogliere dati pubblici.

   Dall'inverno del 2013 questo team di giornalisti svizzeri ha iniziato a sperimentare tecniche di data-driven journalism. Il risultato è stato ottimo, soprattutto per il risultato che questo team ha saputo offrire alla sua redazione.
7. **Best Entry from a Small Newsroom** (miglior progetto per una piccola redazione) – Kiln. Il Kiln è un insieme di strumenti giornalistici, matematici e grafici per trasformare i dati in una storia capace di creare un forte livello di interazione. È un'azienda Britannica che ha aperto la sua banca dati dall'aprile del 2013.
8. **Jurors' Choice** (Progetto scelto dai giurati) – ProPublica, premiata per l'ottimo lavoro svolto quest'anno, migliore tra i progetti che non hanno ricevuto nessun premio.

Se andiamo ad analizzare la geografia delle 75 candidature finaliste potremmo notare una certa disparità, che rispecchia la disparità che esiste nel mondo dell'innovazione digitale.

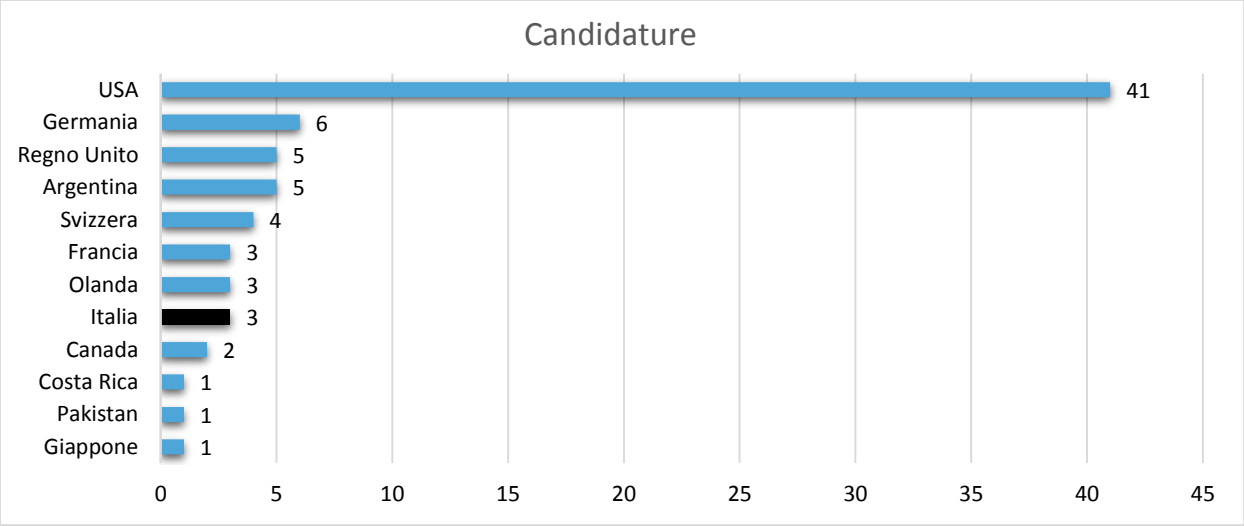


Grafico 1 – DJA Finalist 2014

Dai dati emerge una supremazia imbarazzante da parte degli Stati Uniti con un numero di candidati di gran lunga superiore a tutti gli altri stati. Anche effettuando il confronto tra Stati Uniti e l'intera Unione Europea, la supremazia dei primi rimane inalterata, anzi gli Stati Uniti hanno più candidati rispetto al resto del mondo (41 contro i 34 del resto del mondo). Il maggior numero dei candidati indica non solo lo stato evolutivo nel mondo dell'informatica, ma anche il maggior interesse dei principali giornali americani al data journalism, indicato da molti come il giornalismo del futuro.

3.2 THE MIGRANTS FILES – COME È NATO IL PROGETTO ITALIANO

The Migrants Files è riuscito nell'impresa di vincere un premio per il data journalism, raccontandoci, il dramma dei migranti che cercano di entrare in Europa. Lo scopo di questa inchiesta giornalistica è informare i cittadini dell'Unione Europea su quello che accade nei nostri confini, soprattutto in merito al fatto che il numero di decessi reali è più alto del 50% rispetto al numero dei decessi stimati. Oltre a questo, l'inchiesta mira a mostrarci, attraverso una mappa cliccabile, i luoghi dove queste tragedie sono avvenute, questo per farci capire quali sono le rotte più pericolose per i migranti. Dalla mappa emerge che la rotta più tragica è sicuramente il Mar Mediterraneo, infatti, quest'inchiesta viene pubblicata in Italia dall'Espresso con il titolo "*Migranti, la guerra del mediterraneo*". Nell'articolo, uscito nel marzo 2014, i numeri di queste tragedie vengono comparate ad una vera e propria guerra, sia per le dimensioni che per il numero di decessi, in media più di 1600 l'anno. Il progetto è ancora in corso e punta ad aumentare la quantità di dati includendo anche i numeri delle tragedie precedenti all'anno 2000.

L'idea per questo progetto Europeo nasce in Italia nell'estate del 2012, parte dalla redazione dell'attuale Dataninja, quando ancora non si chiamava in questo modo. Tutto parte dalla raccolta dati di Gabriele Del Grande, *Fortress Europe* (fortezza europea), la più grande e precisa raccolta di dati sul fenomeno dei morti durante le migrazioni in Europa. Da questo elenco più quelli ottenuti dai link alla fonte sono state ottenuti più di 1600 decessi avvenuti negli ultimi 25 anni. Da quella tabella è stata portata avanti l'operazione di analisi dei dati, con l'obiettivo finale di ricostruire i luoghi dei naufragi, con coordinate

geografiche le più precise possibili, quest'analisi dei dati è stata definita dagli stessi redattori di dataninja come la parte più difficile, in quanto i dati sono stati elaborati tutti a mano, sfruttando solo le funzionalità collaborative di strumenti come Google Drive. La prima mappa di questo lungo progetto è stata pubblicata nell'aprile del 2013, sul sito datajournalism.it, con il titolo "*Mar Mediterraneo, tomba di migranti*", che rappresenta una vera e propria applicazione web interattiva che permette di esplorare i dati in due modalità: la prima esplorando una mappa cliccabile e la seconda attraverso un racconto animato nel tempo. La mappa del Mediterraneo si presenta come una mappa con diverse zone d'intensità, nella quale zoomando è possibile vedere gli eventi raggruppati prima per nazione e poi per città più prossime ai luoghi dei naufragi. I cerchi rossi presenti nella mappa indicano i luoghi della tragedia, la cui dimensione è proporzionale alla gravità della tragedia. Nella modalità "Storia" invece si parte dall'autunno del 2009 e con il tempo che scorre vengono mostrati uno dopo l'altro i luoghi e la gravità dei naufragi, con delle bolle rosse che sembrano esplodere lì dove è avvenuta una tragedia. L'animazione si sofferma su alcuni casi particolarmente gravi o rappresentativi, permettendo di leggerne la storia completa prima di proseguire. Ecco un'immagine del progetto.

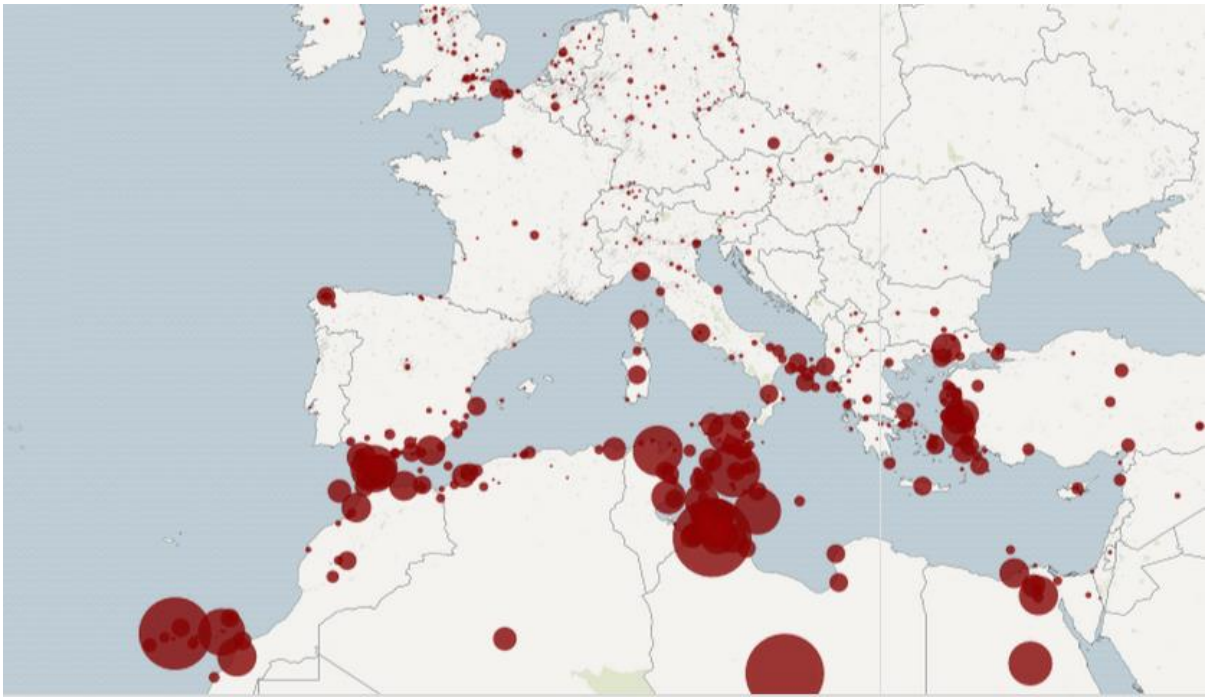


Figura 3 - Mar Mediterraneo, la tomba dei migranti

Quando il progetto non aveva ancora ancora raggiunto notorietà, il 3 ottobre 2013 ci fu una delle tragedie più grandi della storia recente, dove persero la vita 360 persone, tra uomini, donne e bambini. Questo era solo l'ultimo di una serie di naufragi che evidenzia come la vera tomba dei migranti sia principalmente la rotta per Lampedusa. Solo dopo questo evento i redattori di dataninja si resero conto di non poter continuare solo con i dati di Fortress Europe e che da soli non avrebbero potuto arricchire ulteriormente un fenomeno ampio e complesso come quello dell'immigrazione in Europa. Cosi nella seconda metà del 2013 nasce il progetto europeo The Migrants Files, che vede il coinvolgimento di giornalisti provenienti da Svizzera, Svezia, Francia, Germania e Spagna, tutti insieme riuniti in un progetto ambizioso, *riunire e conciliare i maggiori database esistenti sulle vittime delle migrazioni in Europa.* Inizialmente fu integrato, al già citato Fortress Europe, il dataset della *United for Intercultural Action* ed in un secondo momento i dati estratti da *PULS,* un progetto dell'università di Helsinki. Nell'ottobre del 2013 il progetto viene premiato con un finanziamento di 7000 euro, che permettono di portare avanti il lavoro. Una delle maggiori difficoltà incontrata dai giornalisti è stata la ripulitura e la trasformazione dei diversi dataset, dato che per essere uniti è necessario capire a fondo la metodologia che gli autori originari hanno utilizzato, affrontando la grande questione della duplicazione eventi.
Solo alla fine di questo lungo lavoro e grazie alla collaborazione con diverse agenzie d'informazione europee è stato possibile creare una mappa interattiva che raccoglie i dati provenienti da tutta Europa. In seguito, grazie a *Frontex,* agenzia dell'Unione Europea deputata al controllo e alla gestione dei confini, è stato possibile fornire una stima della mortalità nelle zone di frontiera, individuando quelle che sono le zone più calde. L'Italia è interessata da ben due rotte: quella tra Libia e Sicilia e quella tra il nord Africa verso Puglia e Calabria. L'inchiesta finale è stata pubblicata il 31 marzo 2014, in contemporanea in diversi paesi europei, tra cui *El Confidencial* (Spagna), *Neue Zürcher Zeitung* (Germania), *Sydsvenskan* (Svezia) e *Le Monde Diplomatique* (Francia). Il dataset finale è accessibile e consultabile mediante la piattaforma Detective.io, all'indirizzo *http://themigrantsfiles.com*.

3.3 LE ALTRE REALTÀ ITALIANE

Tra i 75 finalisti, oltre al progetto creato dai datanijia, erano presenti altri due progetti italiani, che non sono riusciti a ottenere nessun premio. *Wired.it*, alla seconda partecipazione, dopo aver raggiunto la finale l'anno scorso, è arrivato in finale anche quest'anno nella categoria "Best Story on a single Topic", con un'inchiesta intitolata *Il prezzo della politica italiana: 5 miliardi di euro in 20 anni.*

In questo articolo si mette sotto accusa il costo della politica italiana, in particolare si sottolinea che i soldi che hanno incassato negli ultimi 20 anni sono di oltre 5 miliardi di euro, tra i quali la fetta maggiore è rappresentata dai finanziamenti pubblici ai partiti, erogati sotto forma di rimborsi elettorali. Tutti questi numeri sono stati inseriti in un archivio online e sono liberamente consultabili da tutti i cittadini. Gli stessi dati hanno permesso di creare una infografica dinamica che permette di esplorare tutti i finanziatori privati della politica italiana, grandi e piccoli. Oltre a questo wired ha creato WP (Wired-politcs) un indice che misura la capacità dei partiti di attrarre donazioni. I dati presi in considerazione vanno dal 1992 sino ad oggi, una constatazione interessante è data dal fatto che nel 1993 c'è stato un referendum per l'abolizione dei finanziamenti pubblici ai partiti e che di fatto sarebbero dovuti essere aboliti.

L'altro candidato italiano era *Libero/La Stampa*, candidato per ricevere il premio come miglior raccolta dati per un'organizzazione. L'organizzazione di Libero è arrivata in finale non per un solo progetto ma per ben quattro progetti. In tutti si evidenzia l'ottima infografica presente.

Italia, un territorio fragile – Qui si focalizza l'attenzione sul dissesto idrogeologico al quale sta andando in contro il nostro paese, nel quale si verificano continuamente frane, allagamenti e alluvioni. Le aree ad alta criticità rappresentano il 9.8% della superficie nazionale, dove sorgono 6.250 scuole e 550 ospedali. Le cause di tutto questo vengono attribuite soprattutto alla cementificazione selvaggia: il consumo del suolo è aumentato del 156% dal 1956 ad oggi, causando, negli ultimi cinquant'anni, la morte di quattromila persone.

Perché la Germania guida l'eurozona – si mettono a confronto i principali stati dell'eurozona sotto diversi punti di vista, lavoro, disoccupazione, istruzione, aspettativa di vita, spesa sanitaria ecc. In questi dati emerge la posizione dominante della Germania, usata sempre come metro di paragone per gli altri stati dell'eurozona, dove emerge, più che altro, la drammaticità della situazione greca. Anche qui troviamo un'infografica navigabile, con un'interfaccia molto chiara e completa. I dati esaminati vanno dal 2002 sino al 2012.
Siria: dalla primavera araba alla guerra civile – un'infografica mostra la situazione della guerra in Siria e la sua importanza strategica per la stabilità del rapporto tra occidente e medio oriente. Scorrendo la grafica è possibile vedere i numeri del conflitto, 92.901 le vittime stimate nel periodo di tempo che va dal marzo 2011 sino all'aprile del 2013. Ben più di 1.700.000 i rifugiati, la maggior parte in Turchia. Oltre a questo, sempre scorrendo la grafica è possibile vedere le posizioni dei vari stati internazionali che entrano in gioco nella guerra, dalla presenza dell'unico porto russo sul mediterraneo, alla volontà dei paesi occidentali di bloccare in Siria l'ondata di rivoluzioni partita dal nord africa nel 2010.
Quando l'Italia trema – qui si centra l'attenzione sulla situazione sismica nella quale si ritrova il nostro paese, l'infografica ci permette di vedere che 22 milioni di persone vivono in zone ad elevata pericolosità sismica, con il 43.3% del territorio nazionale ad elevato rischio sismico. Anche qui si cerca di ricostruire il totale dei danni economici prodotti dai terremoti. Viene fatto un paragone tra le situazioni che ci sono tra le diverse regioni italiane, anche se mancano i confronti, che sarebbero molto più significativi, con altri stati mondiali ad alto rischio sismico.

3.4 THE EDITORS LAB

Il "*The Editors Lab*" è un programma che ha come obiettivo principale la crescita e lo sviluppo delle tecnologie usate nell'ambito del giornalismo dei dati e del giornalismo in generale. Il programma, alla seconda edizione, prevede l'organizzazione dei già citati "hack days" da parte delle più importanti redazioni giornalistiche mondiali, trai i quali troviamo il *New York Times, The Guardian, El Pais, Le Parisien, Die Zeit, Il gruppo editoriale l'espresso*. Questi incontri danno la possibilità di sviluppare delle collaborazioni tra

le redazioni, per facilitare la nascita di nuove idee e la creazione di una "*Best practice*" che serva da esempio per tutto il settore del datajournalism.

Quest'anno sono stati organizzati 17 hack days in 17 diverse città, in tutti e cinque i continenti. I vincitori di questi 17 hack days sono stati invitati al Summit finale a Barcellona tenutosi lo scorso giugno dove è stata consegnata la coppa del mondo per la redazione più innovativa. Agli Hack days hanno partecipato più di 170 team, ogni team doveva essere composto da tre persone: un giornalista, un designer grafico e un programmatore. Ogni team doveva rappresentare una società mediatica. I lavori presentati da queste squadre sono stati valutati in base a sette elementi: qualità editoriale, originalità, fonti utilizzate, originalità tecnica del progetto, adeguatezza del tema e facilità d'implementazione.

Il vincitore è stato il team del *New York Times,* con il progetto *"Moments"*, un servizio molto innovativo, che permette di integrare, secondo per secondo, video e testo sullo schermo. Gli utenti possono così guardare un video e leggere un articolo, nel quale le parole coincidono con le immagini nel video. L'utente può naturalmente mandare avanti o portare indietro il video selezionando un punto nel testo.

Al secondo posto si è piazzato il team del *Financial Times,* con un progetto che punta a migliorare e soprattutto velocizzare la visualizzazione del video, permettendo agli utenti di saltare la parte iniziale, ritenuta non interessante, ed andare direttamente ai punti più interessanti del video. Una parte del video viene valutata interessante in base al comportamento che hanno gli utenti che guardano il video. Per i giornalisti questo è uno strumento perfetto perché gli permette di sapere quali sono i punti che sono piaciuti di più agli utenti e a sapere in quali punti la redazione deve migliorare. Gli utenti possono scegliere la parte del video più interessante cliccando sull'immagine che rappresenta il pollice alto o basso.

Al terzo posto troviamo *Wirtualna Polska* con "*Daily Carrot*", il suo team ha prodotto una app per il cellulare che permette agli utenti di creare dei report, in tempo reale, su temi di attualità, permettendo anche di effettuare una ricerca rapida all'interno degli stessi report pubblicati. Come ha detto lo stesso membro di Wirtualna Polska "questo è un ottimo metodo per mettere l'utente di fronte alle notizie".

4 L'IMPORTANZA DEGLI OPEN DATA

Gli open data sono una tipologia di dati liberamente accessibili a tutti, privi di libretto o altre forme di controllo. Anche se il data journalism non si basa esclusivamente sugli open data, essi rappresentano una fonte d'informazione molto importante. In questi anni gli open data hanno permesso al data journalism di crescere, soprattutto perché hanno messo a disposizione dei giornalisti e dei cittadini dati di pubblico interesse, come quelli sulla spesa pubblica. Uno dei problemi principali è rappresentato dal fatto che molti dei dati usati dai giornalisti non sono pubblici e accessibili a tutti e questo ne limita molto la trasparenza. Per capire l'importanza degli open data dobbiamo capirne la filosofia che contempla la condivisione dei dati, i quali possono essere utilizzati e ridistribuiti da chiunque, in modo da permettere un grande passo in avanti dal punto di vista della disponibilità dei dati e soprattutto dal punto di vista della trasparenza, grazie a questo ogni individuo può controllare la veridicità dei dati pubblicati anche da casa sua. Esistono alcuni esempi pratici che hanno dimostrato l'efficacia e l'utilità degli open data anche dal punto di vista economico, come il progetto finlandese "*tax tree*" (l'albero delle tasse) e il Britannico "*Where does my money go?*" (dove vanno i miei soldi?), che permettono di identificare come i soldi delle tasse dei cittadini sono impiegati dal governo. Un esempio interessante è quello del Canada, dove gli open data hanno fatto risparmiare 3.2 miliardi di dollari in un caso di frode fiscale legato alla beneficienza.

4.1 LA CULTURA OPEN DATA

Con "Open data" o "dati aperti" si fa riferimento ad una filosofia e ad una pratica di condivisione di determinati dati e determinate informazioni. Secondo la "open definition" vengono definiti come:

"*I dati che possono essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, soggetti eventualmente alla necessità di citarne la fonte e di condividerli con lo stesso tipo di licenza con cui sono stati originariamente rilasciati*".

La “Full Open Definition” ci spiega in dettaglio cosa questo significhi. Gli aspetti più importanti sono:

- **Disponibilità e accesso:** i dati devono essere disponibili nel loro complesso, per un prezzo non superiore ad un ragionevole costo di riproduzione, preferibilmente mediante scaricamento da Internet. I dati devono essere disponibili in un formato utile e modificabile.
- **Riutilizzo e ridistribuzione:** i dati devono essere forniti a condizioni tali da permetterne il riutilizzo e la ridistribuzione. Ciò comprende la possibilità di combinarli con altre basi di dati.
- **Partecipazione universale**: tutti devono avere la possibilità di accedere, usare, riutilizzare e ridistribuire i dati. Non ci devono essere discriminazioni né di ambito di iniziativa né contro soggetti o gruppi. Ad esempio, la clausola ‘non commerciale’, che vieta l’uso a fini commerciali o restringe l’utilizzo solo per determinati scopi (es. quello educativo) non è ammessa.

Gli open data fanno parte di un movimento più ampio chiamato “*Open Government Data*”, concetto per il quale tutte le attività prodotte dai governi e dalle amministrazioni, debbano essere “aperte” e “trasparenti”. In questa ottica è facile capire l’importanza che possono avere gli open data, l’apertura permetterebbe di sviluppare un dialogo e collaborare con i cittadini, la comunicazione tra le amministrazioni e i cittadini permetterebbe di sviluppare un discorso diretto e partecipativo, quindi un rapporto bidirezionale, con un ruolo importante dei cittadini nei processi decisionali. Mentre la trasparenza, che implica la pubblicazione di tutte le informazioni del settore pubblico (in modo completo, gratuito e accessibile a tutti), permetterebbe ai cittadini di essere in prima fila nel controllo delle attività pubbliche e sapere quello che fa il proprio governo. Il principio è lo stesso che esiste per i software “Open Source”, i cittadini possono lavorare insieme e migliorare diversi elementi della pubblica amministrazione, e di conseguenza risparmiare una grande quantità di denaro pubblico.

Ci sono anche numerose categorie di soggetti e organizzazioni che possono trarre beneficio dalla disponibilità di dati aperti, inclusa la pubblica amministrazione. È già possibile indicare un vasto numero di aree dove i dati pubblici stanno già creando valore:

- Trasparenza e controllo demografico;
- Partecipazione;
- Accrescimento dell'influenza dei cittadini nella discussione pubblica;
- Innovazione;
- Miglioramento dei servizi privati;
- Miglioramento dell'efficienza dei servizi pubblici

Anche dal punto di vista economico gli open data hanno un'enorme importanza. Svariati studi hanno stimato il valore economico degli open data in diverse decine di miliardi di euro ogni anno. Un esempio a riguardo potrebbe essere il sito danese "husetsweb.dk" che aiuta a trovare i modi migliori per risparmiare energia elettrica in casa. Anche Google Translate sfrutta gli open data, usa l'enorme volume di documenti dell'Unione Europea, disponibili in tutte le lingue d'Europa, per allenare gli algoritmi di traduzione automatica. Ci sono numerosi esempi dove gli open data stanno già creando dei vantaggi economici e sociali, ancora non sappiamo quali saranno i possibili utilizzi futuri. Nuove combinazioni di dati possono creare nuova conoscenza e nuove intuizioni. Questo potenziale non sfruttato può essere utilizzato se si riesce a far diventare open data i dati delle pubbliche amministrazioni. Questo è possibile solo se l'apertura è completa e non ci sono limitazioni al riutilizzo dei dati.

4.2 OPEN DATA NEL MONDO

Dalla prima pubblicazione degli open data negli Stati Uniti, nel 2009, ci sono stati notevoli sviluppi, infatti, sempre più stati decidono di mettere a disposizione i loro dati pubblici, anche se, come sostenuto dalle ricerche svolte dalla OKFN (Open Knowledge Foundation), è emerso che lo sviluppo degli open data nel mondo si sta evolvendo in ma-

niera diseguale. Nel sito della OKFN è presente una tabella dove gli stati vengono classificati in base alla completezza degli open data che sono stati resi pubblici dalle rispettive amministrazioni pubbliche. La classificazione viene fatta assegnando un valore da 0 a 100 ad ogni "campo" che viene ritenuto rilevante. I campi rilevanti sono 10:

1. Transport Timetable (Orari dei trasporti)
2. Government Budget (Budget del governo)
3. Government Spending (Spese del governo)
4. Election Result (Risultati delle elezioni)
5. Company Register (Registro imprese)
6. National Map (Mappe nazionali)
7. National Statistics (Statistiche nazionali)
8. Legislation (Legislazione)
9. Postcodes/Zipcodes (codice postale)
10. Emissions of pollutants (emissioni inquinanti)

Per ognuno di questi campi si controllano se i dati pubblicati rispettano i seguenti requisiti:

1. Does the data exist? (Esiste il database?)
2. Is data in digital form? (È in forma digitale?)
3. Publicly available? (È disponibile pubblicamente?)
4. Is the data available for free? (È disponibile gratuitamente?)
5. Is the data available online? (È disponibile online?)
6. Is the data machine readable? (È in un formato leggibile dai computer?)
7. Is available in bulk? (C'è una grande disponibilità?)
8. Is Openly licensed? (Ha una licenza "Open"?)
9. Is the data provided on a timely and up to data basis? ("I dati sono forniti in un periodo tempestivo e su un database"?)

Secondo questi elementi ecco la classifica dei venti paesi più "aperti"

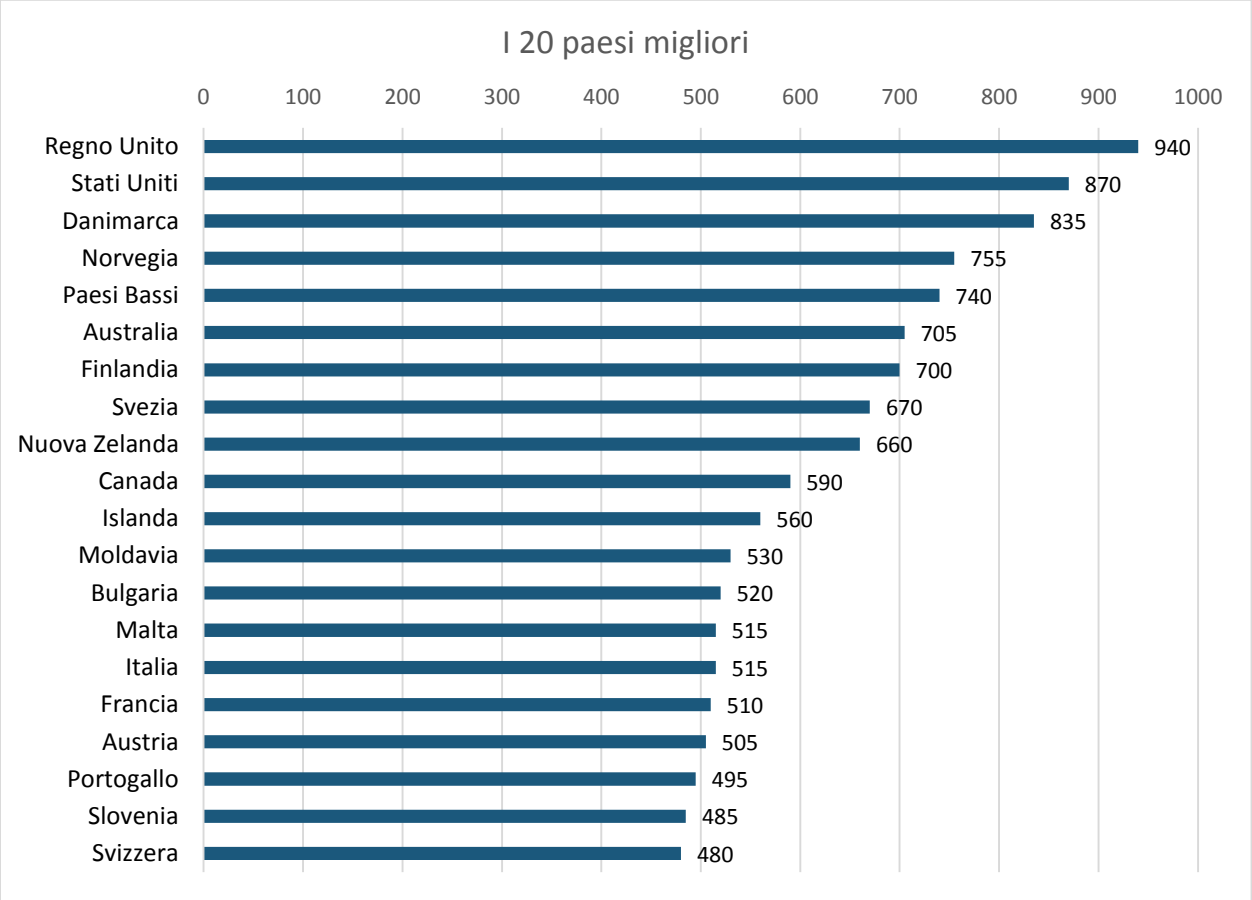


Grafico 2 - Open Knowledge Foundation

Secondo i parametri adottati la nazione che rispetta di più i criteri degli Open data è il Regno Unito con l'ottima valutazione di 940 punti su un massimo di 1000, questo significa che pubblica quasi tutti i dati richiesti dalla Open Knowledge Foundation, fanno molto bene anche Stati Uniti e Danimarca, rispettivamente con 870 e 835 punti. Per trovare l'Italia dobbiamo andare sino alla posizione numero 14 (pari punti con Malta), che con un punteggio di 515[4] non raggiunge nemmeno la sufficienza. Dal grafico è possibile notare che siamo di pochissimo più avanti della Francia, solo 10 punti di differenza, ma anche dopo paesi come Bulgaria e Moldavia.

[4] Aggiornato al 05/01/2015

4.3 *OPEN DATA IN ITALIA*

L'Italia ha iniziato ad interessarsi al data journalism e agli open data in ritardo rispetto agli altri Paesi del mondo, il concetto cominciò a diffondersi solo durante il Festival del giornalismo del 2010. Nel seguente grafico è possibile esaminare la situazione dell'Italia più in dettaglio.

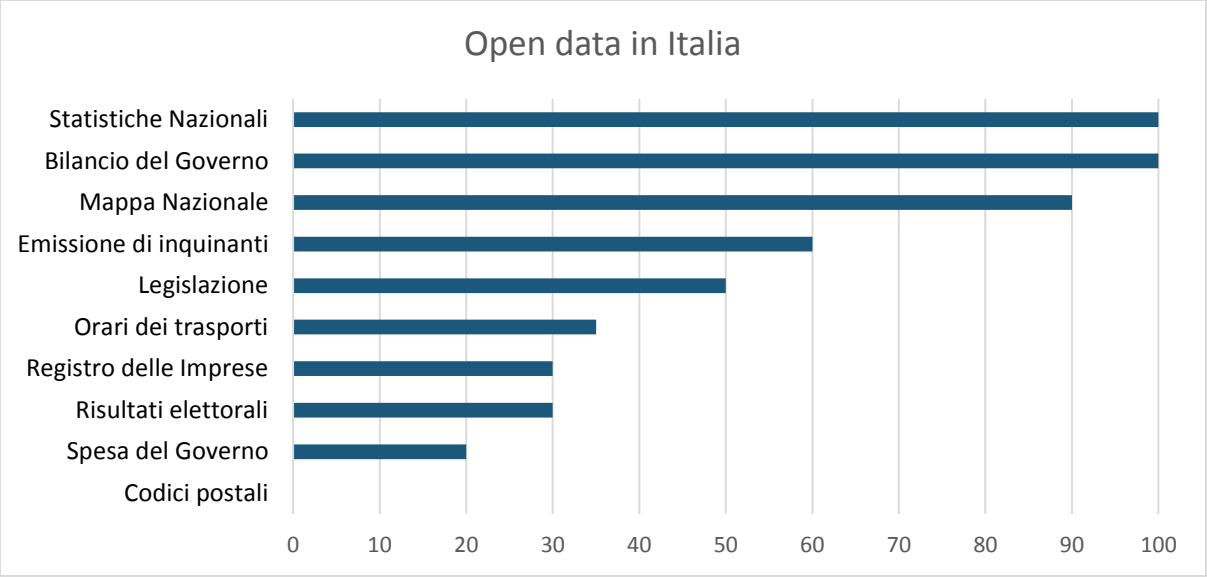


Grafico 3 - Open Knowledge Foundation (Italia)

Si può notare che ci sono 2 settori dove l'Italia ha raggiunto il punteggio massimo di 100 su 100, questi settori rappresentano le statistiche nazionali dell'Istat[5] e il sito della "Ragioneria Generale dello Stato"[6], dove viene pubblicato il bilancio del governo. Ottima anche la valutazione che è stata data riguardo alla mappa nazionale che ha rispettato quasi tutti i requisiti. I risultati peggiori si raggiungono con le voci "codici postali" dove non risulta esserci nessuna voce a riguardo, male anche per i dati riguardanti la "spesa del governo".

4.4 *COME OTTENERE I DATI?*

Spesso uno dei problemi che si trovano davanti gli utenti non è il fatto che i dati non siano presenti nella rete ma, più che altro, il problema è trovarli e dopo averli trovati è

[5] Raggiungibile all'indirizzo http://dati.istat.it
[6] Raggiungibile all'indirizzo http://www.rgs.mef.gov.it/VERSIONE-I/

sempre bene verificare l'attendibilità della fonte. A riguardo però gli esperti che hanno creato il manuale del datajournalism citano degli importanti consigli su come trovare i dati. Un metodo sempre valido per cercare informazioni in rete è usare un motore di ricerca, anche se si consiglia di usare alcuni accorgimenti per facilitare la ricerca delle informazioni. Google, Bing e altri motori di ricerca permettono di personalizzare la ricerca, per esempio si potrebbe restringere la ricerca ai formati di dati che più ci interessano, aggiungendo alla ricerca "filetype:XLS" o "filetype:CSV" se siamo interessati ad un foglio elettronico, se invece siamo interessati ad un database è possibile aggiungere alla ricerca "filetype:MBD","filetype:SQL" o "filetype:DB". Un'altra soluzione può essere quella di inserire come parola chiave il nome di uno dei siti web amministrativi che mettono a disposizione grandi quantità di dati, per esempio aggiungendo alla ricerca "site:data.gov",

Oltre ai siti delle pubbliche amministrazioni negli ultimi anni sono apparsi una grande quantità di siti e portali che trattano grandi quantità di dati, questi sono dei buonissimi posti per trovare le informazioni di cui abbiamo bisogno, tra i più importanti troviamo il sito datahub.io/, gestito dalla già citata Open Knowledge Foundation, che permette di condividere, cercare e riutilizzare fonti di dati aperte. I datablog dei giornali più importanti possono essere una grande fonte di dati, in particolare i siti dei giornali più attenti al data journalism, come il "Guardian World Government Data". Un ottimo modo per ottenere dei database, è quello di cercarli negli appositi fourm, come Get The Data[7] o Quora[8], si può controllare in questi siti se qualcuno ha già richiesto i dati che noi stiamo cercando, queste comunità sono molto utili soprattutto per ricevere indicazioni su dove trovare i dati o anche per ricevere informazioni su come trattare i dati. Tra i più rilevanti troviamo anche dei siti che mettono a disposizione degli indici aggiornati dove è possibile trovare una elenco dei siti dove è possibile scaricare dei dati come *Opendata500.* Un altro esempio è *OpenStreetMap,* nato con l'obbiettivo di creare e rendere disponibili dati cartografici, liberi e gratuiti, dato che la maggior parte delle mappe, che si credono libere, hanno delle restrizioni che ne limitano l'uso agli utenti.

[7] http://getthedata.org/
[8] http://www.quora.com/

Tra i principali portali italiani troviamo *Spaghetti Open Data*, il portale dei *Dati Aperti della PA* e *il portale italiano degli Open Data*.

È possibile controllare in questi siti se qualcuno ha già richiesto i dati che noi stiamo cercando, queste comunità sono molto utili soprattutto per ricevere indicazioni su dove trovare i dati o anche per ricevere informazioni su come trattare i dati. Quando non è possibile trovare i dati di cui si ha bisogno possiamo rivolgerci direttamente alle pubbliche amministrazioni, l'accesso ai dati è regolato dai singoli Stati con norme che variano da nazione a nazione. La tabella seguente mostra alcuni dei più importanti siti che permettono di scaricare i dati pubblici.

Fonti	**Link**
Goverment Data	http://data.gov
Europ's Public Data	http://publicdata.eu/
Datahub	http://datahub.io
The Guardian	http://www.theguardian.com/data
The Upshot (NYT)	http://www.nytimes.com/upshot/
Get the Data	http://getthedata.org
Quora	http://quora.com
Opendata 500	http://www.opendata500.com/
Datamarket	https://datamarket.com/
Spaghetti Open Data	http://www.spaghettiopendata.org/
Dati Aperti della PA	http://www.dati.gov.it/
Il portale Italiano degli Open Data	http://www.datiopen.it/
Open Data Hub	http://www.opendatahub.it/

4.5 COME "APRIRE" I DATI?

La apertura dei dati, non riguarda solo i dati delle amministrazioni pubbliche, bensì tutti i tipi di dati. Nel sito italiano "Opendatahandbook.org" è presente una guida che mira a fornire consigli concreti e dettagliati a colore che possiedono dei dati e che intendano aprirli.

Secondo questa guida ci sono 3 regole principali da seguire:

1. **Scegliere la semplicità.** Non è necessario aprire tutti i dati in una sola volta. Inizialmente va bene aprire anche un solo dataset, o una sua parte. Naturalmente più dataset vengono pubblicati meglio è.
2. **Coinvolgere gli utenti fin dall'inizio.** Cercare sin da subito e spesso il confronto con i potenziali utilizzatori dei dati tra cittadini, imprese o sviluppatori, ciò aumenterà la rilevanza dell'iniziativa durante tutti il suo percorso. È essenziale tener presente che gran parte dei dati non raggiungeranno gli utenti finali direttamente, ma attraverso degli intermediari che probabilmente opereranno delle operazioni sui nostri dati.
3. **Affrontare i timori e le incomprensioni diffuse.** Questo è importante soprattutto se lavori in o con grandi organizzazioni come le istituzioni governative.

4.6 FREEDOM OF INFORMATION (DIRITTO DI ACCESSO ALLE INFORMAZIONI)

Il diritto di accesso agli atti amministrativi è un diritto fondamentale dei cittadini, indicato tra i diritti umani dagli organismi internazionali, sia dall'Onu che dall'Unione Europea. In Italia la legge che si occupa del diritti di accesso alle informazioni è la n° 241/1990. Questa legge permette un accesso più limitato rispetto ad altri paesi in Europa e nel resto del mondo, infatti dice che un cittadino per richiedere l'accesso alle informazioni deve "*avere un interesse diretto, concreto e attuale, corrispondente a una situazione giuridica tutelata*"[9], questo significa che il cittadino ha il diritto di accedere a queste informazioni solo per tutelare il proprio interesse giuridico. Questa norma è stata poi modificata nel 2005, permettendo l'accesso agli individui che dimostrino di rappresentare un interesse pubblico, in questa categoria possono rientrare anche i giornalisti e le associazioni di difesa dei consumatori.

Un confronto tra la nostra legge (241/1990) e quella in vigore negli altri paesi europei e in USA, mostra il ritardo dell'Italia sia dal punto di vista culturale che legislativo. La norma

[9] Ripreso dalla legge 241/1990, http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1990-08-07;241!vig

italiana è infatti l'unica in Europa a concedere l'accesso alle informazioni solo nel caso in cui ci sia un diretto interesse giuridico del cittadino, nel resto d'Europa e negli USA, al contrario, il diritto di accesso è garantito a chiunque, indipendentemente dal suo scopo, diventando un vero e proprio strumento di controllo dell'attività amministrativa. Per questo motivo da qualche tempo in tanti si sono mobilitati per adottare il cosiddetto FOIA (Freedom of Information Act), che obbliga la pubblica amministrazione a rendere pubblici i propri atti e rende possibile a tutti i cittadini controllare le attività amministrative del governo. Come già detto in precedenza questa apertura e trasparenza, in particolare nel Regno Unito e negli Stati Uniti, ha mostrato degli effetti positivi sul funzionamento della pubblica amministrazione, anche per questo, questa legge potrebbe essere utile in uno stato come l'Italia conosciuto per la sua lenta burocrazia.

5 CASI DI STUDIO

Un settore in cui i dati sono sempre stati protagonisti è quello sportivo. Le analisi sportive fanno un abbondante uso di numeri provenienti da svariate statistiche, chiaramente gli elementi presi in considerazione variano da sport a sport. In particolare negli Stati Uniti i telecronisti amano descrivere le partite basandosi sui numeri che sono riusciti a raccogliere, nel prepartita si parla dei numeri relativi alle partite precedenti e nel post-partita si analizza la partita basandosi sulle statistiche raccolte. Negli Stati Uniti i tre sport più seguiti sono, in ordine di preferenza, il football americano, il baseball e il basketball, anche se tra i giovani è in forte ascesa la passione per il calcio, che nel resto del mondo è lo sport più seguito. Attraverso l'analisi di questi dati gli esperti cercano di individuare i "fattori nascosti", che solamente i dati possono rivelare, il tutto per cercare di capire e di prevedere come andranno le partite oppure interi campionati.

Negli Stati Uniti la passione per le statistiche nasce molto tempo fa. Le prime statistiche sportive riguardano il baseball e compaiono negli anni cinquanta sul *JASA (Journal of the American Statistical Association)*. Successivamente negli anni sessanta, settanta e ottanta si diffondono in tutti gli altri sport, football, basketball, golf, tennis, e hockey. Negli anni gli appassionati dei dati aumentano e si arriva negli anni novanta alla creazione del *SIS (Statistics In Sports)*, che sin dalla sua creazione cerca di aprire un dibattito sulle diverse statistiche che vengono usate nei diversi sport, attraverso dei meeting annuali.

5.1 L'ANALISI DEI DATI NELLA NBA

Come abbiamo già detto, le analisi sportive negli Stati Uniti sono sommerse dai dati, il campionato di basket più seguito al mondo, la NBA, non fa eccezione. I numeri raccolti guidano i commenti dei giornalisti, sono molto utili per la valutazione dei singoli giocatori e delle squadre, anche se non tutti credono nell'importanza delle statistiche. Uno dei più grandi cestisti di sempre, *Bill Russel*, è dell'opinione che "*L'unica statistica rilevante è il punteggio finale*". Le statistiche non sono riservate solo agli addetti al lavoro, ma sono

disponibili a tutti sul sito *nba.com* e sul sito dell'emittente televisivo americano *ESPN*. Per ogni giocatore si calcolano i numeri dei minuti giocati, tiri dal campo realizzati, tiri dal campo tentati, percentuale dei tiri realizzati, tiri da 3 tentati, tiri da 3 realizzati, percentuale dei tiri da 3 realizzati, tiri liberi tentati, tiri liberi realizzati, percentuale di tiri liberi realizzati, poi ancora rimbalzi offensivi, rimbalzi difensivi, rimbalzi totali, assist, palle perse, palle rubate, stoppate, punti per partita e plus/minus. Oltre a tutti questi numeri, a partire dal 2010, le emittenti televisive si sono accordate per installare a bordocampo delle speciali telecamere per seguire contemporaneamente tutti i giocatori in campo, permettendoci per esempio di sapere quanti chilometri percorre un giocatore in una partita. Riguardo a questa nuova tecnologia i pareri sembrano essere discordanti, in particolare i giocatori sembrano essere contrari a questo tipo di tecnologie. *Rajon Rondo*, che gioca per i *Boston Celtics*, sostiene infatti che molti di questi numeri non ti permettono di capire quanto un giocatore si impegni, quanto cuore ci metta o quanto giochi duro. D'altra parte, l'allenatore degli stessi Boston Celtis, meglio conosciuto come *Doc. Rivers*, ha accolto la notizia con più entusiasmo, affermando che "*può essere fatto davvero un buon utilizzo di questi dati, possiamo usarli noi e le altre squadre, questi dati dipendono da come un giocatore gioca e da come difende, è possibile scoprire tante cose nascoste*"[10]. L'interesse crescente per le statistiche da parte delle squadre ci viene dimostrato anche dal fatto che sino a dieci anni fa, come sostenuto da alcune fonti della NBA, c'erano forse due squadre che avevano un membro dello staff con il ruolo di analista dei dati, mentre ora, ogni singola squadra ha nel suo organico almeno un membro dello staff che presenta il titolo di analista dei dati. In questo particolare momento è molto diffusa, anche in Italia, l'usanza di valutare il giocatore usando il plus-minus, anche se alcune emittenti televisive stanno cercando di introdurre dei nuovi parametri per misurare il valore dei giocatori, come il *Real plus-minus.*

Plus-minus

Il commentatore sportivo *Dean Oliver*, ci spiega come, più dieci anni fa, l'introduzione di una nuova statistica metrica, il *plus-minus* abbia cambiato il modo di valutare i

[10] Fonte: www.bostonglobe.com [sit_NBA01]

giocatori. Dean afferma, forse con un po' di presunzione, che a suo parere questo nuovo modo di valutare i giocatori ha cambiato il mondo del basket, consentendo di valutare l'impatto che il giocatore ha avuto sulla partita. Questa statistica è stata utilizzata per la prima volta nelle partite di hockey nel 1968, ed è arrivata al basket solo nel 2003. Tuttavia, questo dato si è rivelato più utile nel basket che nell'hockey. Il plus-minus permette di valutare il rendimento dei giocatori basandosi sul conteggio della differenza tra canestri fatti e canestri subiti durante la permanenza del giocatore sul campo. Se una squadra subisce un canestro da 2pt mentre il giocatore è in campo, tutti i giocatori in campo ricevono -2 di plus-minus, invece se la squadra segna un canestro da 2pt tutti i giocatori in campo ricevono +2 di plus-minus, la somma di tutto questo conteggio ci permette di valutare se il rendimento di un determinato giocatore per la sua squadra è stato positivo o negativo. Nel suo libro, *Basketball on paper*, Dean ci parla di come spesso le statistiche che ci vengono mostrate non ci permettono di capire l'andamento della partita e le prestazioni dei singoli giocatori, e si senta quindi la possibilità di introdurre dei nuovi criteri di valutazione.

I nuovi fattori di valutazione

Il plus/minus permette di capire l'impatto che un giocatore ha sulla partita, ma non riesce a valutare il vero lavoro del giocatore sul campo. Si è cercato quindi nel tempo di introdurre dei nuovi metodi di valutazione, tra cui l'*Advanced Plus-minus*, poi il *Regularized Adjusted Plus-minus*. Questi metodi di valutazione non sono stati giudicati abbastanza precisi, cosi, sempre l'emittente televisiva ESPN, ha cercato di introdurre degli altri metodi di valutazione, nell'estate nel 2011 ha introdotto il *Player Efficiency Rating (PER)*, che è un metodo di valutazione che riguarda soprattutto la fase offensiva. L'ultima invenzione di ESPN è il *Real Plus-Minus (RPM)*, che, a detta degli addetti ai lavori, dovrebbe permettere di valutare il vero contributo che viene dato dal giocatore in campo. L'RPM è stato sviluppato da *Jeremias Engelmann*, che fa parte dello staff dei *Phoenix Suns*, in collaborazione con *Steve Ilardi*, professore in psicologia nell'università del Kansas. Questa nuova statistica ci permette di avere una valutazione dei giocatori più precisa, mentre con il classico plus-minus per ogni punto fatto dalla squadra viene attribuito un +1 di plus-

minus a tutti i giocatori in campo, con questa nuova regola si cerca di dare una valutazione più precisa sul lavoro fatto sul campo da ogni giocatore, si valuta innanzitutto la fase offensiva (ORPM) e la fase difensiva (DRPM) e poi effettuando la somma dei due si ottiene il RPM. Il valore attribuito ai giocatori è sempre relativo ai punti segnati dalla sua squadra, solo che il calcolo si basa sui possessi di palla che la squadra effettua con quel determinato giocatore in campo, e non solo sui punti segnati. Questa è solo l'ultima statistica introdotta negli ultimi anni, solo il tempo ci dirà se questa statistica sarà adottata ufficialmente o se passerà nel dimenticatoio. Dato che gli addetti ai lavori stanno già studiando altri metodi per valutare le prestazioni dei giocatori.

5.2 LA PREDIZIONE NEGLI EVENTI SPORTIVI

Predire il futuro è sogno che da sempre attanaglia l'umanità, nel settore sportivo, i giornali che si occupano di data journalism provano a farlo basandosi sui dati raccolti nelle partite precedenti. Le previsioni possono essere fatte letteralmente in un numero infinito di modi, infatti le diverse metodologie scelte portano a diverse previsioni, quindi spesso il problema per chi cerca di fare delle previsioni è capire quali dati possono essere rilevanti e quali no. Le previsioni più affascinanti emergono quando si mescolano diversi fattori. Tuttavia a volte nello sport succedono delle cose che non possono essere previste, tante volte si sono viste delle "corazzate" cadere contro squadre che avevano solamente l'1% di possibilità di vittoria. Questo è il bello dello sport, nulla è ancora scritto.

5.2.1 Previsione nei playoff NBA (2014)

La maggior parte delle previsioni che vengono fatte prima dei playoff NBA sono il frutto di accurate riflessioni personali basate su quello che gli "esperti" del settore hanno visto durante l'anno. Questi esperti si basano in parte sui dati raccolti dalle squadre durante l'anno, ma spesso queste opinioni sono influenzate da sensazioni e da fattori interpersonali. Un metodo molto usato negli Stati Uniti per prevedere il risultato di eventi sportivi è il metodo del *Log5* di *Bill James*. Questo metodo permette di calcolare l'andamento dei playoff basandosi sui risultati reali che le squadre hanno raggiunto nella regular season per

prevedere i risultati che le squadre raggiungeranno nei playoff. Log5 non tiene conto degli infortuni dei giocatori, quindi se un giocatore chiave si infortuna nei playoff la sua squadra avrà le stesse possibilità di vittoria, con o senza di lui. Mettiamo il caso che la previsione riguardi il team *A* contro il team *B,* la probabilità che il team *A* batta il team *B* è uguale a:

$$p_{A,B} = \frac{p_A - p_A \times p_B}{p_A + p_B - 2 \times p_A \times p_B}$$

Dove p_A è il coefficiente di vittorie raggiunte da *A* contro *B* durante l'anno, p_B è il numero di vittorie raggiunte da *B* contro A. Di conseguenza se abbiamo $p_A = 1$ allora il team *A* avra il 100% di chance di vittoria, al contrario, se abbiamo $p_A = 0$, allora sarà il team *B* ad avere il 100% di chance di vittoria. Questa formula è stata ideata per il baseball e Dean Oliver ha adattato questo sistema alla NBA. Qui sotto possiamo vedere le percentuali di vittoria stimate per il primo turno dei playoff 2014.

Conference Est	Pacers	91%	Bulls	67%	Raptors	87%	Heat	94%
	Hawks	9%	Wizards	33%	Nets	18%	Bobcats	6%
Conference Ovest	Spurs	87%	Rockets	70%	Thunder	91%	Clippers	68%
	Mavericks	13%	Blazers	30%	Grizzlies	9%	Warriors	32%

I *Pacers* avevano il 91% di possibilità di passare il turni, mentre gli avversari, gli *Hawks,* avevano solo il 9% di possibilità di passare.

Le otto squadre che nella realtà hanno superato il turno sono in ordine *Pacers, Wizard, Nets, Heat* per la conference dell'Est, e *Spurs, Blazers, Thunder, Clippers* per la conference dell'Ovest. Il Log5 ci ha permesso di indovinare più della metà delle partite del primo turno, sbagliando 3 previsioni su 8, delle quali solamente una era veramente difficile da immaginare, ovvero il passaggio del turno dei Nets dato al 18%. Vediamo quali sono state le previsioni per il secondo turno.

Conference Est	Pacers	74%	Raptors	26%
	Bulls	26%	Heat	74%
Conference Ovest	Spurs	71%	Thunder	52%
	Rockets	29%	Clippers	48%

Anche se nelle previsioni ipotetiche sono presenti delle squadre che in realtà non sono hanno superato il primo turno, riescono a passare il turno tutte le squadre con la probabilità più alta, ovvero *Pacers, Heat, Spurs, Thunder*. In questa situazione il metodo di Bill James ha funzionato in modo efficiente.

Finale di conference dell'est		**Finale di conference dell'ovest**	
Pacers	44%	Spurs	61%
Heat	56%	thunder	39%

Anche in questo caso hanno passato il turno le due squadre favorite, *Heat* e *Spurs*. Che si presentano quindi alla finale.

Finale di Nba	
Spurs	66%
heat	44%

Quindi come aveva previsto il *Log5* i vincitori del campionato NBA del 2014 sono stati i *San Antonio Spurs*. A favore di questo metodo di previsioni possiamo dire che ha una buona percentuale di successo, il problema principale è che questo si basa solamente sui risultati che le squadre hanno ottenuto durante la stagione regolare, quindi, sostanzialmente, si aggiudicherà il titolo la squadra che è andata meglio nella stagione regolare. Ciò chiaramente non è sempre vero, anzi non succede molto spesso, anche perché si dovrebbe tener conto dello stato di forma dei giocatori, degli infortuni e di tanti altri fattori. Tuttavia questo metodo di previsione viene usato spesso e in molti altri sport americani.

5.2.2 I mondiali di calcio 2014

I mondiali di calcio, a livello di audience, sono una delle competizioni sportive più seguite al mondo. Molti data journalist hanno provato a predire il vincitore finale o perlomeno l'andamento nella fase iniziale, si cerca ogni volta di prendere in considerazione più fattori per riuscire ad avere una previsione più precisa. In questa sezione vengono prese in esame due tipi di previsioni: quelle fatte prima dell'inizio del mondiale, che inevitabilmente si basano sui risultati delle partite precedenti ai mondiali e le previsioni che vengono fatte basandosi sulle prime partite del girone le quali dovrebbero permettere di accumulare dei dati più recenti e più significativi.

Previsioni premondiale

FiveThirtyEight ha basato le sue previsioni sull'SPI (Soccer Power Index) di ESPN, un sistema che combina partite e valutazioni dei giocatori per calcolare le prestazioni delle nazionali durante la coppa del mondo. SPI elabora i dati delle partite giocate in un modo complesso: le partite precedenti vengono "pesate" in base all'importanza e al livello dell'avversario, il peso dipende anche dal periodo nel quale sono state giocate, quelle più recenti hanno un peso maggiore, viene inserito un vantaggio per le squadre che giocano in casa e infine vengono valutate le prestazioni dei giocatori con i propri club. FiveFhirtyEight è nota per la sua non trasparenza e questo algoritmo non ci viene spiegato nei dettagli. Un altro lavoro degno di nota è quello di *Andrew Yuan*, che lavora per conto dell'*Economist*, Yuan, nel suo lavoro, mostra nei dettagli qual è stato il suo procedimento, le tecnologie utilizzate e quali dati sono stati presi in considerazione. Nella prima parte, quella della raccolta dati, sono stati presi in considerazione 41.444 partite ufficiali dal 1930 ad oggi, 46.842 posizioni nel ranking FIFA per 215 nazioni dal 1993 sino ad oggi, ed infine ha preso in considerazione la geolocalizzazione dei dati per 838 partite delle 32 squadre qualificate al mondiale. Nella seconda fase ha calcolato la percentuale di vittorie rapportandolo alla differenza di rango, usando come unità di misura il ranking FIFA, con i risultati ottenuti ha fatto una stima dei diversi scenari possibili e ha stimato le probabilità di vittoria di ogni squadra. Un'altra importante analisi è quella che riguarda un report della *Goldman Sachs*,

il loro modello si basa su un modello stocastico che cerca di prevedere i possibili risultati delle 64 partite in programma nel mondiale, prendendo in considerazione una serie di eventi a partire dall'anno 1960. Anche in questo caso viene preso in esame il ranking FIFA, il numero di goal medio segnato nelle ultime cinque partite, il numero di goal subiti nelle ultime 5 partite, un indicatore dei risultati delle squadre nelle precedenti edizioni della coppa del mondo, e un'altra variabile per vedere in che modo una squadra gioca le partite in casa. Quest'ultimo metodo non include nessun valore individuale dei giocatori, vengono presi in considerazione solo i risultati a livello di squadra. Nella tabella sottostante possiamo vedere il confronto tra le diverse previsioni.

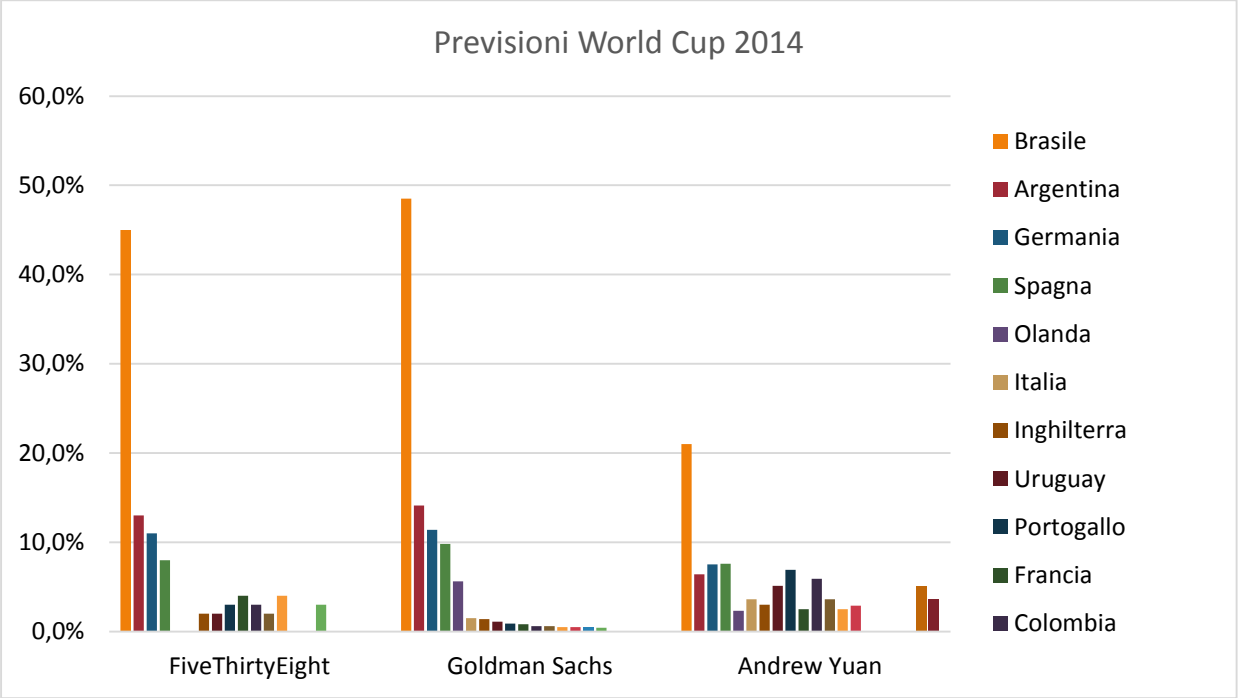


Grafico 4 - Previsioni premondiale 2014

In tutte e tre le previsioni premondiali prese in esame il Brasile risulta essere di gran lunga il favorito, mentre tra le altre squadre c'è un maggiore equilibrio e anche un'alternanza tra le posizioni previste. La seconda squadra ad essere favorita in due schemi è l'Argentina e in uno la Spagna, mentre la terza è sempre la Germania, per la quarta posizione troviamo due volte la Spagna e una volta l'Argentina, le altre previsioni invece sono diverse e meno uniformi tra i tre schemi. In realtà queste previsioni non si sono avvicinate tantissimo alla realtà dei risultati, dato che la Spagna è stata eliminata nella fase iniziale, la Germania,

prevista come terza invece ha vinto dominando in quasi tutte le partite disputate, soprattutto con il Brasile, con un imbarazzante 7 a 1. L'unica che ha mantenuto le attese è stata l'Argentina che si è piazzata seconda. Questo dimostra come sia difficile fare delle previsioni per tornei così brevi e dove ci sono di mezzo delle partite ad eliminazione diretta.

Previsioni dopo la prima fase

Queste previsioni vengono fatte alla fine delle fasi a gironi, prima dell'inizio della fase ad eliminazione diretta, rispetto alle previsioni premondiali è possibile inserire dei dati più recenti, dato che questi si riferiscono alle tre partite giocate da ogni squadra nei gironi. Il sito *Bigdatatales.com,* gestito da *Luca Pappalardo* e *Paolo Cintia*, presenta un modello nel quale ogni team viene valutato in base al movimento della palla e dei giocatori, in modo da analizzare in che modo ogni squadra "attacca la palla"[11]. In questo modello, nelle partite iniziali del mondiale, è emerso che le squadre vincitrici erano state quelle che possedevano una maggiore eterogeneità e imprevedibilità nei passaggi. Basandosi su questi risultati sono state previste tutte le restanti partite, dagli ottavi di finale sino alla finale, aggiornando i valori ottenuti dopo ogni turno giocato.

Il sistema usato si è dimostrato ottimo riuscendo a prevedere quasi tutti i risultati, l'unica "eccezione" è data dalla partita dei quarti di finale *Francia – Germania*, nella quale è stata la Germania a vincere nonostante la Francia possedesse una più alta eterogeneità.

L'elemento peculiare di questo modello è che i risultati si aggiornano dopo ogni partita, e si riferiscono solo alla "*Attualità*" e non fanno riferimento al "curriculum" delle squadre che si presentano al mondiale. La tabella sottostante rappresenta i punteggi assegnati alle squadre prima della semifinali.

Semifinali			
Brasile	2.333	Olanda	1.604
Germania	2.918	Argentina	2.938

Tabella 3 -Proiezioni semifinali del mondiali, bigdatatales.com

[11] Fonte: http://bigdatatales.com/2014/07/12/taca-la-bala-says-the-wizard-a-trip-into-the-world-cup-2014/

I dati mostrano che Germania e Argentina sono le squadre che hanno maggiori probabilità di arrivare in finale, riuscendo a prevedere quello che è accaduto nella realtà.

Finale	
Germania	3.716
Argentina	3.233

Tabella 4 - Proiezioni finali dei mondiali - bigdatatales.com

La tabella 4 mostra la proiezione della finale, in questo caso la larga e agevole vittoria per 7 -1 della Germania contro il Brasile nella partita di semifinale, ha permesso alla Germania di diventare la favorita per la vittoria finale del torneo. Mentre prima delle semifinali, come per altro è possibile constatare nella tabella 3, la squadra con il maggior punteggio era l'Argentina. Questo modello ha dimostrato di possedere ottime potenzialità predittive dato che è riuscito a prevedere quasi tutto l'andamento del mondiale dai sedicesimi di finale in poi. Il suo vantaggio rispetto ai modelli visti in precedenza è che valuta le prestazioni delle squadre rispetto alla competizione attuale e i valori attribuiti alle squadre cambiano dopo ogni partita giocata. Grazie a questo, il modello, è stato in grado di notare le non brillanti prestazioni del Brasile (la favorita) nella fase a gironi iniziale, e di prevederne l'uscita alle semifinali.

5.3 I DATI DIFFICILI, IL FATTORE CAMPO

Quando si parla di eventi sportivi si tende a considerare sempre il fattore campo, ovvero il fatto di giocare in casa o in trasferta, infatti, i dati ci dicono che le squadre che giocano in casa riescono spesso ad avere dei risultati più favorevoli rispetto alle squadre che giocano fuoricasa. Non si è certi di quali siano le vere ragioni di questi vantaggi, forse può influire l'apporto del pubblico, il miglior feeling con il terreno di gioco, l'effetto della stanchezza che deriva dai viaggi fatti per raggiungere il campo avversario, o l'effetto placebo, in alcuni casi anche il fattore climatico può influire sull'andamento della partita, tut-

tavia è indubbio che ci siano dei vantaggi per chi gioca in casa. Il problema dei data journalist è riuscire a capire in che modo e con quali percentuali, il fatto di giocare in casa influenzi il risultato. Anche le federazioni sportive riconoscono questo vantaggio, ad esempio, nel calcio, nei tornei dove sono previste delle partite di "andata" e "ritorno", in caso di parità di differenza reti, si attribuisce più valore ai goal realizzati in trasferta. Nei Playoff della NBA, alla squadra vincitrice della stagione regolare e anche alle squadre che raggiungono un buon piazzamento, viene affidato il vantaggio del fattore campo rispetto alle squadre che si sono piazzate peggio nella stagione regolare. Un'indagine a riguardo ci arriva dal dipartimento di statistiche dell'università della Pennsylvania, il quale prova a quantificare in percentuali il vantaggio del fattore campo nei quattro maggiori sport americani. I dati sono stati presi dal 2001 al 2006 per il Basket, dal 2001 al 2005 per il football, dal 1998 al 2003 per l'hockey e dal 1991 al 2002 per il baseball.

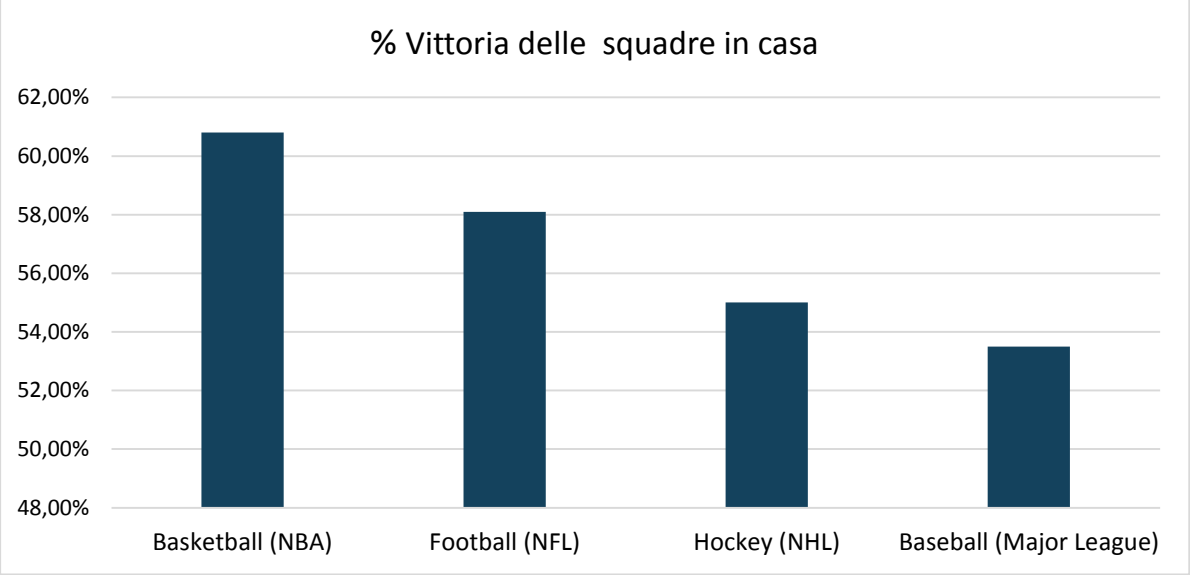


Grafico 5 - Percentuale di vittorie casalinge nei principali campionati sportivi americani

La tabella oltre a mostrarci quello che già sappiano, cioè che c'è un vantaggio per le squadre che giocano in casa, ci mostra che c'è anche un differenza tra i diversi sport. Il basket tra i quattro maggiori sport statunitensi è quello che ha una percentuale più alta, infatti, il 60.8% delle partite giocate vengono vinte dalle squadre che giocano in casa, mentre lo

sport dove è meno influente il fattore campo è il baseball, dove le squadre di casa hanno vinto il 53.5% delle partite giocate.

Il giornale online *Bleacherreport*, specializzato in data journalism approfondisce in un articolo l'importanza del fattore campo, analizzando i dati che vanno dal 2003 al 2011. I risultati mostrano che le percentuali dei dati raccolti sono praticamente identiche alle precedenti, con una percentuale di vittorie pari al 60%. Quando un club gioca "in casa", per ogni partita, il numero delle partite perse diminuisce del 3.1%, il punteggio realizzato aumenta del 3.4%, i contropiedi aumentano del 12.7% e il numero di falli commessi diminuisce del 4.7%, il tutto rispetto a quando si gioca fuori casa. Per cercare di capire quali sono le origini di questi vantaggi sono stati esaminati degli altri fattori e si è arrivati alla conclusione che sono due i fattori che favoriscono le squadre che giocano in casa, la prima è il fattore arbitrale, infatti si crede che il rumore prodotto dai tifosi ad ogni chiamata arbitrale porti il giudice di gare a favorire inconsciamente le squadre di casa, infatti, sempre negli anni tra il 2003 e il 2011 i dati ci dicono che l'arbitro ha fischiato una media di 22.15 falli a partita a favore delle squadre in casa e 21.13 falli di media a favore delle squadre fuori-casa. L'altro fattore che influenza le squadre che giocano in casa è l'effetto placebo, quindi le squadre ottengono migliori risultati in casa solamente perché sono convinte che giocare in casa apporti dei benefici. Anche se esistono delle persone che non credono ad un vantaggio creato dal fatto di giocare in casa i dati dimostrano che questo esiste senza dubbio, più che altro il problema è riuscire a quantificare questo vantaggio.

6 LA CRISI ECONOMICA E IL DECLINO DEL CALCIO ITALIANO

Il calcio italiano sta attraversando un periodo non esaltante, dal 2010 le squadre italiane non sono state all'altezza della propria tradizione calcistica a differenza dei risultati ottenuti nei decenni precedenti, in particolare negli anni 90, nei quali dominava nelle principali coppe europee. In questo periodo c'è stato un vero e proprio crollo dal punto di vista dei risultati, della considerazione e degli investimenti. In questo lasso di tempo si sono verificati alcuni episodi che potrebbero aver influito sull'evoluzione del calcio italiano. L'episodio probabilmente più rilevante è stata la crisi economica che in Europa ha avuto il suo momento peggiore nel 2009, i dati mostrano l'esistenza di alcuni aspetti in comune tra le due crisi che non sono stati osservati negli altri principali campionati europei. È importante notare in che modo, dal momento della crisi, il calcio italiano ha subito un calo dal punto di vista degli investimenti non riscontrato negli altri campionati che, paragonati a quello italiano, sembrano non aver risentito della crisi economica e hanno continuato a investire. I dati inoltre mostrano in che modo i migliori giocatori europei stanno preferendo lasciare il campionato italiano in favore di altre destinazioni.

6.1 GLI EFFETTI DELLA CRISI ECONOMICA

In questa sezione vengono presi in esame gli andamenti economici dei cinque stati che tradizionalmente investono più soldi nel calcio, il campionato italiano, spagnolo, inglese, tedesco e francese. Per esaminare l'andamento economico degli stati sono stati presi in esame i dati sul PIL pubblicati dall'istituto finanziario "World Bank", dal 1999 al 2013. Anche se con qualche differenza le economie degli stati esaminati presentano un andamento simile, con molti aspetti in comune, visibili sul grafico nella pagina seguente.

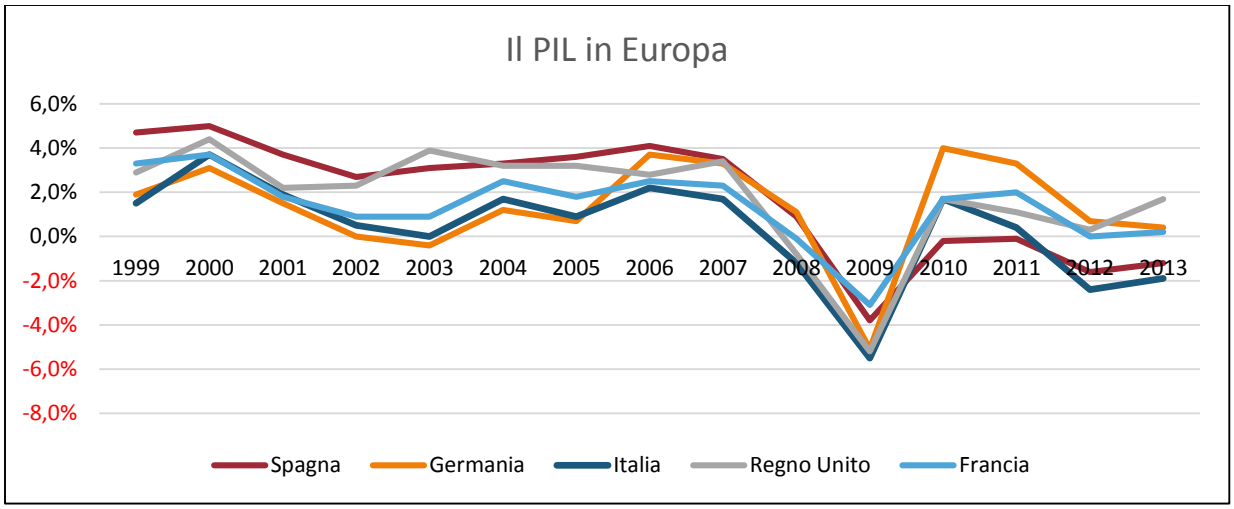


Grafico 6 – Andamento del PIL in Europa (1999-2014) - http://data.worldbank.org/

Dal 1999 al 2007 c'è stato un periodo molto positivo nel quale tutti gli stati esaminati non sono entrati in recessione ma l'aspetto più evidente del grafico è il momento della crisi economica, che ha avuto il suo periodo peggiore nel 2009, nel quale tutti gli stati si sono trovati in recessione. A partire dal 2010 Germania, Regno Unito e Francia sono riuscite a uscire dalla recessione e a riprendere la crescita economica, mentre la Spagna e L'Italia (dopo un 2010 positivo) sono ancora in recessione.

6.2 LA SERIE A È STATA COLPITA MAGGIORMENTE DALLA CRISI ECONOMICA?

Dopo aver dato uno sguardo all'andamento del PIL analizziamo i dati economici che riguardano le società di calcio, per vedere in che modo le economie dei principali club europei hanno risentito della crisi economica. Calcolando il valore medio delle principali squadre dei cinque più importanti campionati nazionali europei è possibile vedere in che modo le economie di queste squadre siano state influenzate dalla crisi. In questa analisi il valore di una squadra è dato dalla somma del valore di mercato dei suoi giocatori. Se una squadra compra dei giocatori il suo valore aumenta, al contrario, se la squadra vende un giocatore, il valore della squadra diminuisce. Usare come metodo di riferimento il valore della rosa di una squadra permette di capire quanto i club stanno investendo e rappresenta un chiaro indice delle condizioni economiche delle aziende che gestiscono i club.

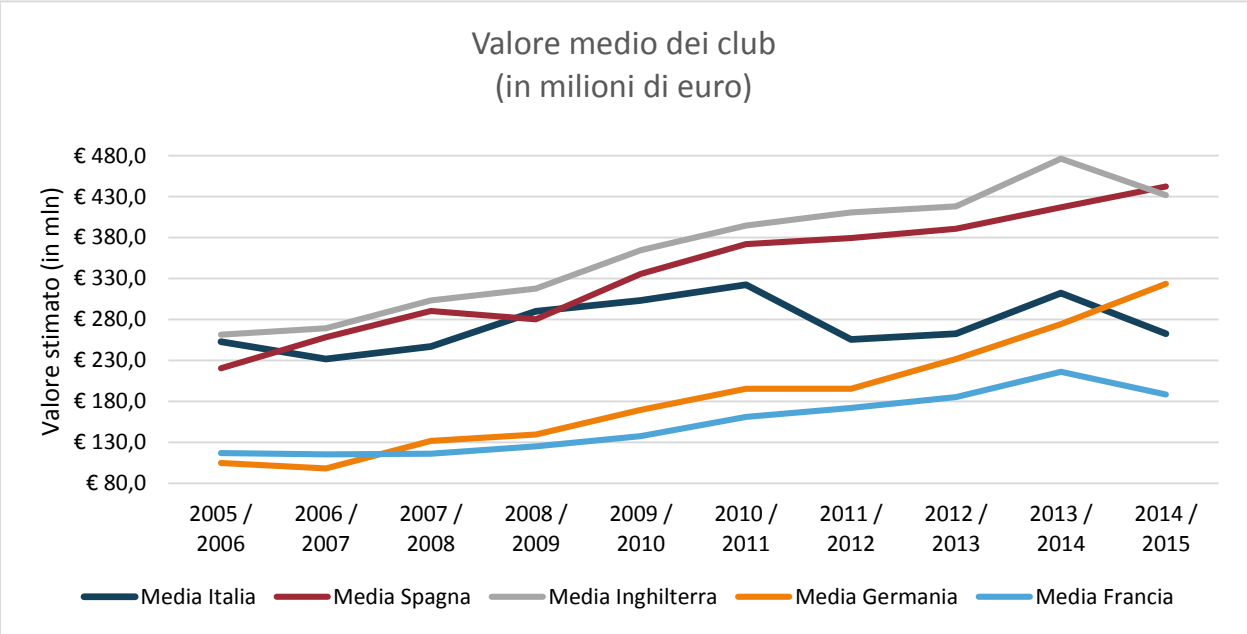


Grafico 7 - Valore medio delle principali squadre dei 5 migliori campionati europei - Transfermarkt.it

Il grafico rappresenta i valori economici dal 2005 al 2014 delle principali squadre dei cinque campionati nazionali più importanti, la serie A italiana, la Liga spagnola, la Bundesliga tedesca, la Premier League inglese e la Ligue 1 Francese. Ogni anno, per ogni campionato è stato calcolato il valore medio delle quattro società più valutate. Il grafico mostra che nel 2005 era possibile suddividere i valore delle squadre in due gruppi, uno più ricco formato da Italia, Spagna, Inghilterra e uno meno ricco formato da Germania e Francia. Tenendo conto che il 2009 è l'anno in cui la crisi economica è stata più acuta, è particolarmente interessante focalizzare l'attenzione su quello che succede a partire dalla stagione 2009/2010, in questa stagione il divario tra il valore economico delle squadre italiane con quelle inglesi e spagnole inizia ad aumentare, nei cinque anni successivi l'andamento degli investimenti delle squadre italiane è molto altalenante, a differenza degli altri campionati nei quali gli investimenti nel complesso continuano ad aumentare e come conseguenza possiamo notare che nella stagione 2014/2015 gli investimenti delle squadre tedesche superano, gli investimenti delle squadre italiane. Per rendere ancora meglio l'idea possiamo analizzare la variazione dei valori medi delle principali squadre dal 2005 a oggi. Il valore

medio delle squadre tedesche è aumentato del 208,49%, passando da 104.8 a 232 milioni di euro. La Spagna ha incrementato il suo valore del 100.64%, passando da 220 mln a 442 mln. Anche l'Inghilterra e la Francia hanno aumentato il valore delle loro squadre rispettivamente del 65.20% e il 61.27%. L'Italia è la nazione che ha registrato il minore aumento. In 9 anni il valore delle squadre italiane è passato dai 252.8 mln di euro, della stagione 2005/2006, ai 262.6 mln di euro della stagione in corso (2014/2015), registrando solo un leggero aumento del 3.9%. Questo è il dato più significativo, che dimostra che il campionato italiano ha subito maggiormente gli effetti della crisi economica.

6.3 CHE RELAZIONE C'È TRA IL VALORE ECONOMICO E IL SUCCESSO FINALE?

In questo paragrafo si vuole osservare la relazione presente tra il valore economico della squadre e le relative possibilità di raggiungere il successo finale nella competizione europea più importante, la Champions League. Il calo del valore delle squadre italiane si riflette anche sui risultati. Un metodo per valutare le prestazioni dei club a livello europeo è il ranking UEFA. Attualmente nelle prime quattro posizioni sono presenti, nello stesso ordine, le quattro nazioni con il valore economico medio più alto. Prima la Spagna, seconda l'Inghilterra, terza la Germania e quarta l'Italia. Questo indica che le squadre che spendono di più ottengono dei risultati migliori rispetto alle altre. A vincere la Champions League però non è sempre la squadra che spende più soldi, ma quella che li spende meglio. Nella tabella sottostante sono presenti i vincitori e i finalisti della Champions League dal 2006 sino al 2014, con un indice che ci mostra il valori dei club in quel determinato anno.

Albo d'oro Champions League – dal 2006 al 2014				
	I vincitori	**Valore**	**Finalisti**	**Valore**
2005 / 2006	Barcellona	6°	Arsenal	8°
2006 / 2007	Milan	5°	Liverpool	7°
2007 / 2008	Manchester United	4°	Chelsea	1°
2008 / 2009	Barcellona	2°	Manchester United	4°
2009 / 2010	Inter	4°	Bayern Monaco	10°
2010 / 2011	Barcellona	1°	Manchester United	5°
2011 / 2012	Chelsea	3°	Bayern Monaco	6°
2012 / 2013	Bayern Monaco	6°	Borussia Dortmund	14°
2013 / 2014	Real Madrid	2°	Atletico Madrid	12°

Tabella 5 - Vincitori e finalisti della Champions League

Possiamo notare che negli ultimi otto anni le squadre che avevano il valore più alto sono riuscite a vincere solamente una volta e ad arrivare in finale solo due volte. Osservando i dati raccolti non è possibile affermare che la squadra con il valore più alto abbia la più alta probabilità di vincere il torneo. Basandoci sui dati possiamo però affermare che per vincere il torneo una squadra debba collocarsi tra le sei squadre più valutate. Dato il numero ristretto di campioni esaminati possiamo attribuire alle sei squadre più valutate all'incirca le stesse probabilità di vittoria. Tra queste sei squadre più valutate ci sono infatti degli altri fattori non economici che entrano in gioco, come l'affiatamento dei giocatori, la completezza della squadra in ogni reparto, l'esperienza nell'affrontare questo tipo di competizioni. A conclusione possiamo affermare che il fatto di investire una grande quantità di denaro è una condizione necessaria ma non sufficiente per vincere la Champions League.

6.4 LA FUGA DEI MIGLIORI GIOCATORI

In il calo degli investimenti nel calcio ha causato la fuga dei migliori giocatori, mentre in passato è stato il nostro campionato ad ospitare tanti dei migliori giocatori ed erano le nostre squadre a spendere per comprare i giocatori, ora dobbiamo stare a guardare le

altre squadre che spendono per rubare i migliori giocatori dal campionato italiano. Il grafico sottostante mostra l'andamento con il quale i giocatori più valutati hanno cambiato le loro destinazioni, preferendo gli altri campionati europei a quello italiano. I dati si riferiscono al periodo di tempo che va dalla stagione 2005/2006 alla stagione 2014/2015.

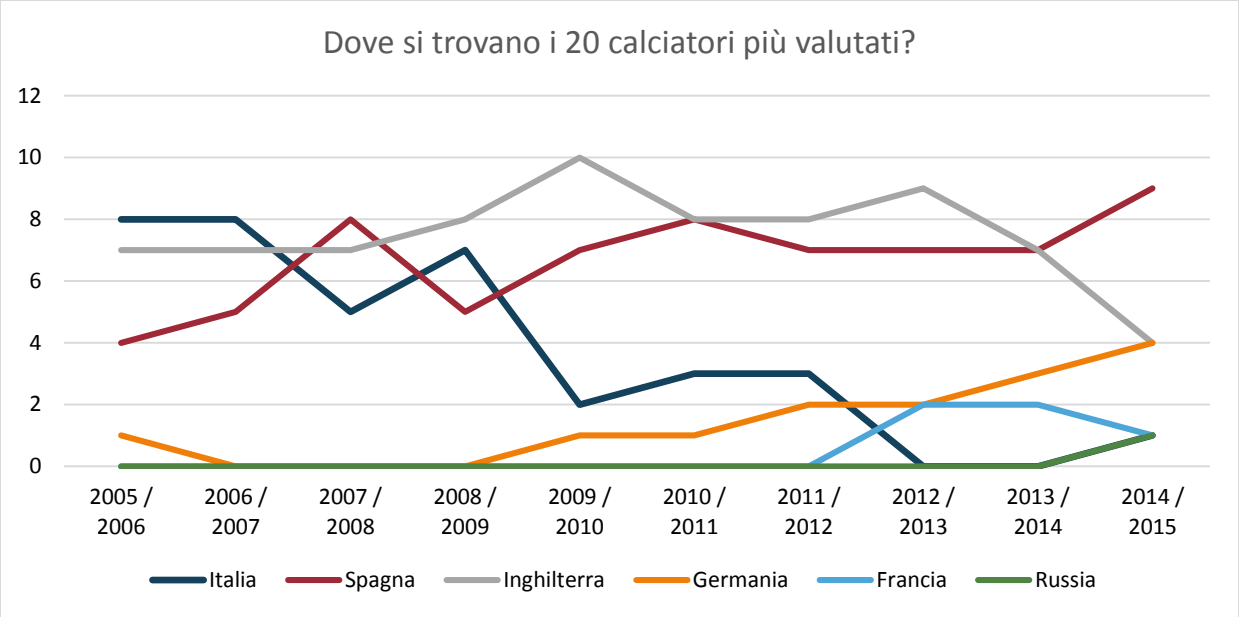


Grafico 8 - posizionamento dei 20 calciatori più valutati

Nella tabella sottostante è possibile vedere i dati più in dettaglio.

Dove si trovano i 20 calciatori più valutati?						
	Italia	Spagna	Inghilterra	Germania	Francia	Russia
2005 / 2006	8	4	7	1	0	0
2006 / 2007	8	5	7	0	0	0
2007 / 2008	5	8	7	0	0	0
2008 / 2009	7	5	8	0	0	0
2009 / 2010	2	7	10	1	0	0
2010 / 2011	3	8	8	1	0	0
2011 / 2012	3	7	8	2	0	0
2012 / 2013	0	7	9	2	2	0
2013 / 2014	0	7	7	3	2	0
2014 / 2015	1	9	4	4	1	1

Tabella 6 - Geoposizionamento dei giocatori più valutati

Osservando Il grafico è possibile notare in che modo, nel campionato italiano, il numero di giocatori della "Top 20" sia progressivamente diminuito, arrivando anche a toccare un deludente zero. Sia nella stagione 2005/2006 che nel 2006/2007, il campionato italiano era quello con il maggior numeri di giocatori tra i 20 più valutati, solo l'Inghilterra si avvicinava. Il dominio italiano finisce nella stagione 2007/2008, dove è la Spagna a ospitare i giocatori più valutati, anche se l'Italia non è molto distante. Il vero calo si nota tra la stagione 2008/2009 e la stagione 2009/2010, che tra l'altro sono gli anni peggiori della crisi economica, l'Italia passa dal possedere sette giocatori tra la top 20 a possederne solo due. Il 2009 è l'anno delle cessioni eccellenti di *Kakà* e *Ibrahimovic*, entrambi vanno in direzione della Spagna. Da quel momento in poi l'Italia non riesce più a rialzarsi, gli investimenti sono sempre più bassi. Il momento meno glorioso della serie A arriva negli anni tra il 2012 e la prima parte del 2014, nel quale non è presente nessun calciatore tra i 20 più valutati e, in questi anni perde anche il terzo posto nel ranking Uefa, a favore della Bundesliga tedesca. Nella stagione in corso, 2014/2015, c'è solamente un giocatore tra i 20 più valutati al mondo, è il francese *Pogba*. I dati della tabella mostrano una certa analogia con il grafico 6, entrambi mostrano la costante crescita della Germania, che nell'ultima stagione presa in esame, ha lo stesso numero di giocatori più valutati dell'Inghilterra. Nelle ultime tre stagioni prese in esame si può notare anche un maggiore decentramento e l'ingresso in questa speciale classifica di altri due campionati, quello francese e quello russo.

6.5 CONCLUSIONI

A dispetto del fatto che la crisi economica ha colpito tutti gli stati europei esaminati, è evidente che tutti i principali campionati, tranne quello italiano, hanno incrementato gli investimenti sul calcio, in particolare la Germania, che oramai sembra un'eccellenza in tutti i settori e non sembra che ci possano essere dei rallentamenti alla sua crescita. A livello economico l'Italia viene spesso paragonata alla Spagna, che, tra gli stati esaminati, sono gli unici due ad essere ancora in recessione, eppure, come abbiamo già visto, il valore delle principali squadre nazionali spagnole è praticamente raddoppiato dal 2005 ad oggi.

I dati quindi mostrano che il campionato italiano non è riuscito a reggere in maniera soddisfacente la crisi economica, a differenza degli altri stati europei. Il legame della crisi del calcio italiano con la crisi economica è netto, anche se probabilmente esistono altri fattori che hanno contribuito ad aumentare il divario con gli altri campionati nazionali. Ad esempio lo scandalo "Calciopoli"[12] del 2006, che ha causato un notevole calo di interesse da parte dei tifosi italiani e la retrocessione di una delle squadre storiche del campionato italiano, la Juventus. Un altro elemento che potrebbe aver influito con il calo del calcio italiano potrebbe essere l'elevata burocrazia e l'alta tassazione italiana, che spesso scoraggiano gli investitori stranieri, che preferiscono molto di più investire nella Premier League Inglese, in questo momento solo due squadre della serie A appartengono a dei proprietari non italiani[13], nella Premier League, i proprietari stranieri sono ben 11 su venti[14].

[12] Il termine indica uno scandalo che ha investito il calcio italiano nel 2006, nel quale sono state coinvolte alcune tra le più importanti società professionistiche italiane,

[13] http://www.ilpost.it/2013/10/15/proprietari-squadre-serie-a/

[14] http://www.theguardian.com/football/2013/nov/19/premier-league-english-richard-scudamore

7 APPROFONDIMENTI SULL'ANALISI DELLA SERIE A

7.1 PERCHÉ USARE IL VALORE DELLE SQUADRE?

Analizzare il valore delle squadre è un ottimo indice per valutare se le squadre hanno deciso di investire, di conseguenza questo permette di analizzare in che modo i proprietari delle squadre di calcio hanno reagito alla crisi economica. Il valore di una squadra aumenta quando si acquistano dei nuovi giocatori o quando dei giocatori fanno delle prestazioni migliori rispetto all'anno precedente, al contrario, il valore diminuisce quando la squadra cede un giocatore, quando un giocatore si infortuna, quando calano le sue prestazioni oppure il giocatore supera una certa età.

Il 2013 è stato un anno dove sono stati registrati degli eventi molto importanti che riguardano il valore delle squadre e dei giocatori. In quest'anno si è registrato un aumento record nel valore medio delle squadre europee esaminate, l'aumento medio è stato di 41.75 mln. La squadra che da una stagione all'altra ha aumentato di più il suo valore è stata il *Chelsea*, sempre nel 2013 il suo valore è passato da 408.50 a 582.98 milioni di euro. Il suo valore è aumentato grazie all'acquisto di diversi calciatori, con una spesa totale di 130 milioni, all'arrivo di alcuni giocatori in prestito, all'aumento del valore di alcuni giocatori già presenti all'interno della sua rosa e all'acquisto di importanti giocatori a parametro zero. Sempre nel 2013 anche in Italia si è registrato un aumento di valore record, il valore della Roma è aumentato di 79 mln, passando 135.9 a 214.55 milioni di euro, grazie all'acquisto di giocatori per un totale di 73.5 mln e all'arrivo di alcuni giocatori in prestito. Nello stesso anno è stato anche effettuato l'acquisto più costoso della storia, *Gareth Bale*, ceduto dal *Tottenham Hotspur* al *Real Madrid* per 94 milioni di euro.

	Valore medio dei club europei	**Valore del Chelsea**	**Valore della Roma**	**Acquisto più costoso**
2012	297.6 mln	408.50 mln	135.90 mln	Thiago Silva 42 mln
2013	339.4 mln	508.98 mln	214.55 mln	Gareth Bale, 94 mln
±	+41.75 mln	+174.48 mln	+78.68 mln	

Tabella 7 - I cambiamenti importanti del 2013

Tra i valori presi in esame c'è una particolarità che riguarda l'aumento del valore medio delle squadre esaminate, i due maggiori aumenti sono stati registrati nelle stagioni precedenti al mondiale, ovvero le stagioni 2009/2010 e 2013/2014.

2006/07	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/2015
3.2 mln	23.3 mln	12.8 mln	31.5 mln	27.0 mln	-6.4 mln	14.9 mln	41.7 mln	-6.7 mln

Tabella 8 - Aumento del valore medio delle squadre europee rispetto all'anno precedente

Basandoci sui dati possiamo supporre che il valore dei giocatori aumenti nelle stagioni precedenti ai mondiali. Durante questo periodo, infatti, aumenta il numero dei tifosi che si interessano al calcio, questo causerebbe un effetto a catena nel quale l'aumento dei tifosi causa una aumento dei profitti derivanti dagli sponsor che a sua volta fa sì che le squadre possano spendere più soldi per acquistare i giocatori, facendo così lievitare il prezzo di molti giocatori. Se questa condizione è vera il prossimo grande aumento, superiore ai 41.7 mln del 2013/2014, si potrà notare solo dalla stagione 2017/2018, che precede i prossimi mondiali che si terranno in Russia.

7.2 ELENCO DELLE SQUADRE ESAMINATE

Le squadre esaminate fanno parte degli attuali cinque campionati nazionali europei che investono più soldi nel calcio. Per ogni campionato sono state prese in considerazione le quattro squadre più valutate anno per anno, dalla stagione 2005/2006 alla stagione 2014/2015. Dato che le valutazioni delle squadre cambiano da un anno all'altro, nell'intero arco temporale sono state prese in esame più di quattro squadre per ogni nazione:

- Cinque per l'Italia: Juventus, Milan, Inter, Roma e Napoli. Delle quali solo la Juventus è sempre rientrata tra le quattro più valutate, anche nell'anno trascorso in serie B nella stagione 2006/2007.
- Cinque per la Spagna: Real Madrid, Barcellona, Atletico Madrid, Valencia e Villareal. Di queste solamente il Real Madrid e il Barcellona sono risultate tra quattro più valutate per tutti gli anni presi in esame.

- Sei squadre per l'Inghilterra: Manchester United, Chelsea, Manchester City, Liverpool, Arsenal, Tottenham. Tra queste solamente il Manchester United e il Chelsea sono risultate sempre tra le quattro squadre inglesi più valutate.
- Sette per la Germania: Bayern di Monaco, Borussia Dortmund, Schalke 04, Bayer Leverkusen, Wolfsburg, Stoccarda, Amburgo. Tra le quali solo il Bayern di Monaco è sempre stato tra le migliori quattro stagionali.
- Sette per la Francia: Paris Saint Germain, Monaco, Olympique Marsiglia, Olympique Lione, Bordeaux, Lille, Saint-Etienne. Nessuna di queste squadre è stata sempre presente tra le quattro squadre nazionali più valutate per tutti gli anni.

7.3 RELAZIONE TRA INVESTIMENTO E RISULTATI

Uno dei fattori più importanti per valutare l'andamento delle squadre è il Ranking Uefa, che assegna dei punti alle squadre che partecipano alle competizioni europee, i punti vengono assegnati ogni qual volta le squadre ottengono dei risultati positivi.
In questa sottosezione analizzo la relazione tra i valori e i risultati delle squadre usando il Ranking Uefa come metodo di valutazione dei risultati. Il grafico sottostante mostra i valori delle principali squadre spagnole, inglesi e italiane che rispettivamente in questo momento occupano il primo, il secondo e il quarto posto del Ranking Uefa

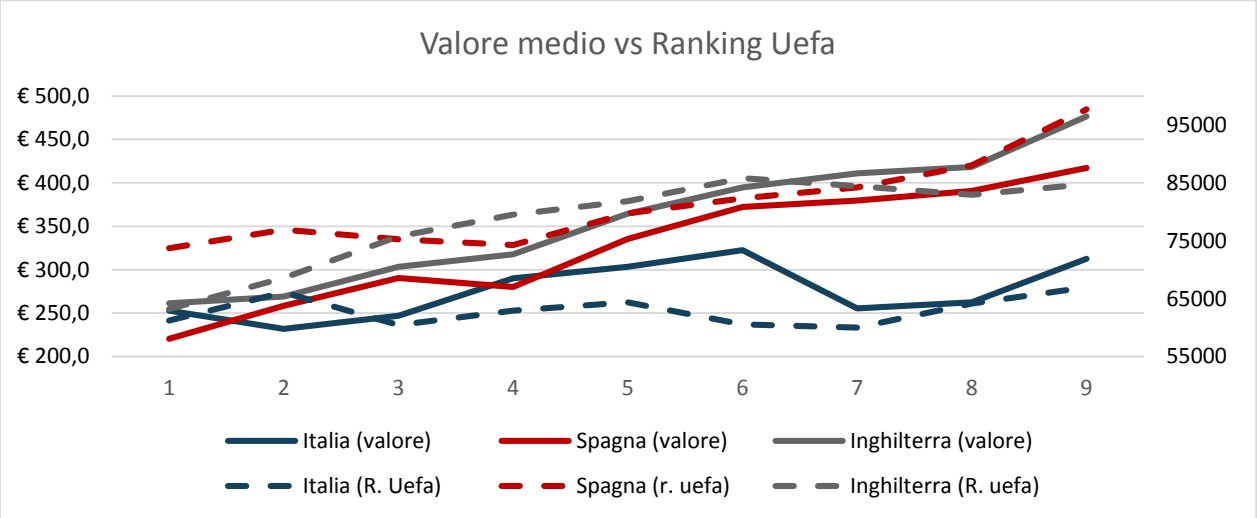


Grafico 9 - Relazione tra il valore medio delle squadre e posizionamento nel R. Uefa

Il grafico evidenzia un andamento simile tra il valore economico e i punti ottenuti nel ranking Uefa, lasciando sottintendere che esiste un legame tra i due dati. In particolare il legame è più evidente se osserviamo i valori relativi alla Spagna (in rosso), ne quale le due linee seguono un andamento molto simile. Il legame è meno evidente se osserviamo i valori dell'Italia (in blu), ne quale il dato che rappresenta il valore delle squadre ha un andamento più altalenante rispetto al valore che rappresenta il Ranking Uefa.

BIBLIOGRAFIA

Briggs Mark (2007) - *Journalism 2.0. How to Survive and Thrive*, ed. Jan Schaffer. 182 pp.

Bolter J. D. & Grusin R. (1999) – *Remediation: Undertanding New Media,* The MIT Press, Boston, 282 pp.

C. P. Scott (1921) – *A Hundred Years* – The Guardian, Manchester.

Crambers Lawrence (1989) - "*Plea for Recognition of Scientific Character of Journalism"*, Journalism Educator, 46-49 pp.

Durrel Huff (2009) – *Mentire con le statistiche,* Monti & Aambrosini editori, Trento, 207 pp.

Gray J, Chambers L., Bounegru L. (2012) - *The Data Journalism Handbook: how Journalistis Can Use Data to Impreve the news*, O'Reilly Media, 242 pp.

Howard Alexander Benjamin (2014) – *The Art and Science of Data-Driven Journalism*, Columbia Journalism School, New York, 144 pp.

Jim Albert, Jay Bennett, James J. Cochran (2005) - *Anthology of Statistics in Sports*, Society for Industrial and Applied Mathematics, 333 pp.

McLuhan Marshall (1967) – *Gli strumenti del comunicare,* Il Saggiatore, Milano, 332 pp.

Meyer Philip (1991) - *The New Precision Journalism*, Indiana University Press, Indianapolis, 273 pp.

Nate Silver (2012) – *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*,Penguin Group, 544 pp.

Peter O'Donoghue (2012) – *Statistics for Sports and Exercise Studies: an introduction*, Routledge, New York, 416 pp.

Robert P. Shumaker, Osama K. Solieman, Hsinchun Chen (2010) – *Sport Data Mining*, Springer US, 157 pp.

Saul L. Miller (2013) – *Perchè i team vincono. Le 9 chiavi del successo nel mondo degli affari dello sport e non solo,* Libreria dello sport, 220 pp.

Stephen Quinn (2002) – *Knowledge Management in the Digital Newsroom*, Focal Press (Taylor & Francis Group), Woburn (Massachesetts), 196pp.

SITOGRAFIA

10 tools that can help data journalist do better work, http://www.poynter.org/how-tos/digital-strategies/147736/10-tools-for-the-data-journalists-tool-belt/, modificato il 10.10.2011, consultato il 30.06.2014.

2014 NBA playoff predictions: Odds favor a Heat vs. Spurs Finals rematch, http://www.sbnation.com/2014/4/17/5620074/nba-playoffs-2014-predictions-odds-heat-spurs, modificato il 17.04.2014, consultato il 10.09.2014

About the Upshot, http://www.nytimes.com/2014/04/23/upshot/navigate-news-with-the-upshot.html?rref=upshot&abt=0002&abg=1, consultato il 24.11.2014

Data Journalism, primo piano, http://www.lastampa.it/medialab/datajournalism/primo-piano, consultato il 24.06.2014

E-book: favorevoli o contrari, http://daily.wired.it/aconfronto/cultura/confronto-sugli-ebook-favorevoli-o-contrari.html, modificato il 28.08.2010, consultato il 20.05.2014.

Evalueting individual player performance indexes in basketball, http://www.stata.com/meeting/italy11/abstracts/italy11_capelli.pdf, consultato 09.09.2014

FiveThirtyEight (old posts on New York Times), http://fivethirtyeight.blogs.nytimes.com/, consultato il 01.09.2014.

FiveThirtyEight's World Cup Predictions, http://fivethirtyeight.com/interactives/world-cup/, modificato il 9.06.2014, consultato il 05.09.2014

Fortress Europe (Database delle persone morte nel mediterraneo), http://fortresseurope.blogspot.it/, modificato il 14.06.2014, consultato il 23.06.2014

GDP growth (annual %), http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG, visitato il 13.10.2014

GEN - Global Editors Network (Official site), http://www.globaleditorsnetwork.org/, consultato il 23.06.2014.

Giornalismo di precisione, http://it.wikipedia.org/wiki/Giornalismo_di_precisione, modificato il 31.03.2014, consultato il 30.05.2014.

Global Editors Network, http://en.wikipedia.org/wiki/Global_Editors_Network, modificato il 14.05.2014, visitato il 23.06.2014.

How numbers have changed the NBA, http://espn.go.com/nba/story/_/id/9980160/nba-how-analytics-movement-evolved-nba, modificato il 15.11.2013, consultato il 09.09.2014.

How Important Is Home-Court Advantage in the NBA, http://bleacherreport.com/articles/1520496-how-important-is-home-court-advantage-in-the-nba, modificato il 08.02.2013, visitato il 04.08.2013

How Much of the NBA Home Court Advantage Is Explained by Rest?, http://www.amstat.org/chapters/boston/nessis07/presentation_material/Dylan_Small.pdf, consultato il 04.08.2013.

I proprietari della seriea A, http://www.ilpost.it/2013/10/15/proprietari-squadre-serie-a/, pubblicato il 15.10.2013, consultato il 20.10.2014.

Il manuale degli Open Data, http://opendatahandbook.org/it/what-is-open-data/, consultato il 10.06.2014

Il New York Times lancia The Upshot, il data journalism come approfondimento, *Il sole 24 ore*, http://www.ilsole24ore.com/art/tecnologie/2014-04-22/il-new-york-times-lancia-the-upshot-data-journalism-come-approfondimento-103102.shtml?uuid=AB2kZuCB, consultato il 29.08.2014

Indice degli open data, https://index.okfn.org, consultato il 13.06.2014

Legge (241/1990), http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1990-08-07;241!vig, consultato il 14.06.2014

Libro cartaceo o ebook? Risponde Umberto Eco, http://www.libriantichionline.com/bibliofilia/libro_cartaceo_o_ebook_risponde_umberto_eco, consultato il 20.05.2014.

Manual de periodismo de datos, http://interactivos.lanacion.com.ar/manual-data/, consultato il 03.06.2014

Mar mediterraneo, tomba di migranti (l'inchiesta), http://stories.dataninja.it/themigrantsfiles/inchiesta/, consultato il 23.06.2014

Open Data Day a Bologna, http://www.dirittodisapere.it/2013/03/05/open-data-day-a-bologna-come-e-andata/, modificato il 20.03.2013, consultato il 14.06.20144

Picking World Cup Winners? After 12 games, FIFA rankings beating eminent thinkers http://www.sportingintelligence.com/2014/06/16/picking-world-cup-winners-after-12-games-fifa-rankings-beating-some-eminent-thinkers-160601/, modificato il 16.06.2014, consultato il 05.09.2014

Precision Journalism and Narrative Journalism http://www.nieman.harvard.edu/reports/article-online-exclusive/100044/Precision-Journalism-and-Narrative-Journalism-Toward-a-Unified-Field-Theory.aspx, modificato il 02.11.2011, consultato il 30.05.2014.

Premier League still 'quintessentially English', says Richard Scudamore, http://www.theguardian.com/football/2013/nov/19/premier-league-english-richard-scudamore, pubblicato il 19.11.2013, consultato il 20.10.2014.

ProPublica, http://en.wikipedia.org/wiki/ProPublica, modificato il 14.05.2014, consultato il 01.09.2014

Ragioneria Generale dello Stato, http://www.rgs.mef.gov.it/, consultato il 13.06.2014

Setting the Record Straight on Hydraulic Fracturing, http://www.propublica.org/article/setting-the-record-straight-on-hydraulic-fracturing-090112, modificato il 12.01.2009, consultato il 01.09.2014.

“Taca la bala” says the wizard: a trip into the world cup 2014, http://bigdata-tales.com/2014/07/12/taca-la-bala-says-the-wizard-a-trip-into-the-world-cup-2014/, modificato il 08.07.2014, consultato il 23.10.2014.

The Color of Money, http://powerreporting.com/color/, consultato il 20.08.2014.

The Detroit riots of 1967 hold some lessons for the UK, http://www.theguardian.com/uk/2011/sep/05/detroit-riots-1967-lessons-uk, modificato il 5.09.2011, consultato il 20.08.2014.

The Migrants’ Files, http://themigrantsfiles.com, consultato il 24.06.2014.

The next big thing: the real plus-minus, http://espn.go.com/nba/story/_/id/10740818/introducing-real-plus-minus, modificato il 07.04.2014, consultato il 10.09.2014.

Web Scraping, http://it.wikipedia.org/wiki/Web_scraping, modificato il 03.05.2014, consultato il 23.10.2014.

What the Fox Knows, http://fivethirtyeight.com/features/what-the-fox-knows/, modificato il 17.03.2014, consultato il 28.04.2014.

What Went Wrong, Miami Herald, December 20.1992, Daniel X. O'Neil, https://www.flickr.com/photos/juggernautco/2844893922/in/set-72157607210036175/, consultato il 26.08.2014.

[WEB_SPI] Soccer Power Index Explained, http://www.espnfc.com/story/1873765, modificato il 11.06.2014, consultato il 02.09.2014

[WEB_Yuan] 2014 Fifa World Cup Brazil predictions (Andrew Yuan), http://andrew-yuan.github.io/EDAV-project.html , modificato il, consultato il 05.09.2014

[WEB_Goldman] The World Cup and Economics 2014 (Goldman Sachs)http://www.goldmansachs.com/our-thinking/outlook/world-cup-and-economics-2014-folder/world-cup-economics-report.pdf, modificato il, consultato il 05.09.2014

[sit_Russell01] Bill Russel Quotes, https://www.goodreads.com/author/quotes/75414.Bill_Russell, consultato il 08.09.2014

[sit_NBA01] New age of NBA analitycs: Advantage or overload?, http://www.bostonglobe.com/sports/2014/03/29/new-age-nba-analytics-advantage-overload/1gAim4yKYXGUQ2CTAe7iCO/story.html, consultato il 09.09.2014.